PLANT GENOME SEQUENCES AND USES THEREOF

This application claims priority under 35 U.S.C §119(e) of U.S. Provisional Applications Nos. 60/108,420, filed November 16, 1998; and 60/120,645, filed February 18, 1999; and under 35 U.S.C. §120 of U.S. Application Serial No. 09/443,025 filed November 12, 1999, the disclosures of which applications are incorporated herein by reference in their entirety.

FIELD OF THE INVENTION

The present invention is in the field of plant biochemistry and genetics. More specifically the invention relates to nucleic acid sequences from plant cells, in particular, genomic DNA sequences from *Arabidopsis thaliana* plants. The invention encompasses nucleic acid molecules present in non-coding regions as well as nucleic acid molecules that encode proteins and fragments of proteins. In addition, the invention also encompasses proteins and fragments of proteins so encoded and antibodies capable of binding these proteins or fragments. The invention also relates to methods of using the nucleic acid molecules, proteins and fragments of proteins, and antibodies, for example for genome mapping, gene identification and analysis, plant breeding, preparation of constructs for use in plant gene expression, and transgenic plants.

BACKGROUND OF THE INVENTION

I. *Arabidopsis thaliana*

Arabidopsis thaliana (*Arabidopsis*) belongs to the *Brassicaceae* plant family -- a commercially important plant family. The identification in *Arabidopsis* of biological agents such as plant promoters, open reading frames, plant gene intron regions, plant gene intron/exon junctions, regulatory elements, genetic and physical markers, and proteins is important in the development of nutritionally enhanced or agriculturally enhanced crops. Such agents are useful in, for example, marker development, genetic mapping or linkage analysis, marker assisted breeding, physical genome mapping, transgenic crop production,

crop monitoring diagnostics, antibody production and gene modification. Such agents can also have pharmaceutical or nutriceutical applications.

Arabidopsis is a small plant in the mustard family and is widely used as a model organism for basic and applied research in the biology of flowering plants. *Arabidopsis* is a model system for plant genomic research in part due to its small and well characterized genome, which has been estimated to be comprised of approximately 20,000 to 25,000 genes. The genome is estimated to have a haploid content of around 100Mb which is present on five chromosomes. Reported partial sequence analysis has provided information on genome features such as gene density and gene structure (Settles and Byrne, *Genome Research 8:*83-85 (1998), the entirety of which is herein incorporated by reference). Based on reports from the European Union Sequencing Consortium, the average gene density is one gene every approximately 4.8kb.

Other important characteristics that make *Arabidopsis* a useful test system include its rapid life-cycle, small size, which allows for controlled growth in restricted space, its prolific seed production, the availability of characterized mutants and the existence of a reliable transformation system.

The value derived from the genome sequence information of *Arabidopsis* is not limited to *Arabidopsis* genetics and biochemistry. *Arabidopsis* belongs to the same plant family, *Brassicaceae*, as oilseed *Brassica* varieties which are a source of edible and industrial vegetable oils. A number of important food crop species, are also members of this plant family, including species from *Brassica* (cabbage, cauliflower, broccoli, kohlrabi, turnips, Brussels sprouts), *Raphanus* (radish) and *Rorippa* (watercress). Other commercially relevant *Brassicacae* products include condiments from *Brassica* (mustard) and *Armoracia* (horse-radish), and ornamentals from about 50 genera, including *Arabis*, *Erysimum* (*Cheiranthus*), *Hesperis*, *Iberis*, *Lobularia*, *Lunaria* and *Matthiola*. Agents and genome sequences of *Arabidopsis* will find particular applications in these closely related plant species by the discovery of syntenic relationships that can be used to identify

regions of interest for genetic modification of valuable crop species. Moreover, *Arabidopsis* exhibits some degree of conserved gene order with the genomes of maize and rice.

II. Sequence Comparisons

Genome sequence information from *Arabidopsis* allow comparisons of *Arabidopsis* sequences with other *Arabidopsis* sequences as well as with those of other flowering plant genome sequences, particularly crop plant species and also with genome sequences and gene sequences from other organisms, including bacteria, humans and yeast. Such information provides valuable insights into the translation of plant genetic information into a flowering plant and also reveals genetic differences involved in the differentiation of the plant kingdom. In addition, genome sequencing and mapping provides increased opportunities for identification and isolation of agents associated with plant traits, as well as insight into mechanisms of genome interactions.

A characteristic feature of a DNA sequence is that it can be compared with other DNA sequences. Sequence comparisons can be undertaken by determining the similarity of the test or query sequence with sequences in publicly available or propriety databases ("similarity analysis") or by searching for certain motifs ("intrinsic sequence analysis") (*e.g.*, *cis* elements) (Coulson, *Trends in Biotechnology 12:*76-80 (1994), the entirety of which is herein incorporated by reference; Birren *et al., Genome Analysis 1*:543-559 (1997), the entirety of which is herein incorporated by reference).

Similarity analysis includes database search and alignment. Examples of public databases include the DNA Database of Japan (DDBJ)(http://www.ddbj.nig.ac.jp/); Genebank (http://www.ncbi.nlm.nih.gov/web/Genbank/Index.htlm); and the European Molecular Biology Laboratory Nucleic Acid Sequence Database (EMBL) (http://www.ebi.ac.uk/ebi_docs/embl_db.html). A number of different search algorithms have been developed, one example of which are the suite of programs referred to as BLAST programs. There are five implementations of BLAST, three designed for

nucleotide sequences queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN) (Coulson, *Trends in Biotechnology 12:*76-80 (1994); Birren *et al., Genome Analysis 1:*543-559 (1997)).

BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and searches them against a nucleotide sequence database. BLASTN was designed for speed, not maximum sensitivity, and may not find distantly related coding sequences. BLASTX takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six translations against a protein sequence database. BLASTX is useful for sensitive analysis of preliminary (single-pass) sequence data and is tolerant of sequencing errors (Gish and States, *Nature Genetics 3:*266-272 (1993), the entirety of which is herein incorporated by reference). BLASTN and BLASTX may be used in concert for analyzing sequence data (Coulson, *Trends in Biotechnology 12:*76-80 (1994); Birren *et al., Genome Analysis 1:*543-559 (1997).

Given a coding nucleotide sequence and the protein it encodes, it is often preferable to use the protein as the query sequence to search a database because of the greatly increased sensitivity to detect more subtle relationships. This is due to the larger alphabet of proteins (20 amino acids) compared with the alphabet of nucleic acid sequences (4 bases), where it is far easier to obtain a match by chance. In addition, with nucleotide alignments, only a match (positive score) or a mismatch (negative score) is obtained, but with proteins, the presence of conservative amino acid substitutions can be taken into account. Here, a mismatch may yield a positive score if the non-identical residue has physical/chemical properties similar to the one it replaced. Various scoring matrices are used to supply the substitution scores of all possible amino acid pairs. A general purpose scoring system is the BLOSUM62 matrix (Henikoff and Henikoff, *Proteins 17:*49-61 (1993), the entirety of which is herein incorporated by reference), which is currently the default choice for BLAST programs. BLOSUM62 is tailored for

alignments of moderately diverged sequences and thus may not yield the best results under all conditions. Altschul, *J. Mol. Biol. 36:*290-300 (1993), the entirety of which is herein incorporated by reference, uses a combination of three matrices to cover all contingencies. This may improve sensitivity, but at the expense of slower searches. In practice, a single BLOSUM62 matrix is often used but others (PAM40 and PAM250) may be attempted when additional analysis is necessary. Low PAM matrices are directed at detecting very strong but localized sequence similarities, whereas high PAM matrices are directed at detecting long but weak alignments between very distantly related sequences.

Homologues in other organisms are available that can be used for comparative sequence analysis. Multiple alignments are performed to study similarities and differences in a group of related sequences. CLUSTAL W is a multiple sequence alignment package available that performs progressive multiple sequence alignments based on the method of Feng and Doolittle, *J. Mol. Evol.* 25:351-360 (1987), the entirety of which is herein incorporated by reference. Each pair of sequences is aligned and the distance between each pair is calculated; from this distance matrix, a guide tree is calculated, and all of the sequences are progressively aligned based on this tree. A feature of the program is its sensitivity to the effect of gaps on the alignment; gap penalties are varied to encourage the insertion of gaps in probable loop regions instead of in the middle of structured regions. Users can specify gap penalties, choose between a number of scoring matrices, or supply their own scoring matrix for both the pairwise alignments and the multiple alignments. CLUSTAL W for UNIX and VMS systems is available at: ftp.ebi.ac.uk. Another program is MACAW (Schuler *et al., Proteins, Struct. Func. Genet, 9:*180-190 (1991), the entirety of which is herein incorporated by reference, for which both Macintosh and Microsoft Windows versions are available. MACAW uses a graphical interface, provides a choice of several alignment algorithms, and is available by anonymous ftp at: ncbi.nlm.nih.gov (directory/pub/macaw).

Sequence motifs are derived from multiple alignments and can be used to examine individual sequences or an entire database for subtle patterns. With motifs, it is sometimes possible to detect distant relationships that may not be demonstrable based on comparisons of primary sequences alone. Currently, the largest collection of sequence motifs in the world is PROSITE (Bairoch and Bucher, *Nucleic Acid Research 22:*3583-3589 (1994), the entirety of which is herein incorporated by reference). PROSITE may be accessed via either the ExPASy server on the World Wide Web or anonymous ftp site. Many commercial sequence analysis packages also provide search programs that use PROSITE data.

A resource for searching protein motifs is the BLOCKS E-mail server developed by S. Henikoff, *Trends Biochem Sci. 18:*267-268 (1993), the entirety of which is herein incorporated by reference; Henikoff and Henikoff, *Nucleic Acid Research 19:*6565-6572 (1991), the entirety of which is herein incorporated by reference; Henikoff and Henikoff, *Proteins 17:*49-61 (1993). BLOCKS searches a protein or nucleotide sequence against a database of protein motifs or "blocks." Blocks are defined as short, ungapped multiple alignments that represent highly conserved protein patterns. The blocks themselves are derived from entries in PROSITE as well as other sources. Either a protein or nucleotide query can be submitted to the BLOCKS server; if a nucleotide sequence is submitted, the sequence is translated in all six reading frames and motifs are sought in these conceptual translations. Once the search is completed, the server will return a ranked list of significant matches, along with an alignment of the query sequence to the matched BLOCKS entries.

Conserved protein domains can be represented by two-dimensional matrices, which measure either the frequency or probability of the occurrences of each amino acid residue and deletions or insertions in each position of the domain. This type of model, when used to search against protein databases, is sensitive and usually yields more accurate results than simple motif searches. Two popular implementations of this

approach are profile searches (such as GCG program ProfileSearch) and Hidden Markov Models (HMMs)(Krough *et al., J. Mol. Biol. 235*:1501-1531 (1994); Eddy, *Current Opinion in Structural Biology 6*:361-365 (1996), both of which are herein incorporated by reference in their entirety). In both cases, a large number of common protein domains have been converted into profiles, as present in the PROSITE library, or HHM models, as in the Pfam protein domain library (Sonnhammer *et al., Proteins 28*:405-420 (1997), the entirety of which is herein incorporated by reference). Pfam contains more than 500 HMM models for enzymes, transcription factors, signal transduction molecules, and structural proteins. Protein databases can be queried with these profiles or HMM models, which will identify proteins containing the domain of interest. For example, HMMSW or HMMFS, two programs in a public domain package called HMMER (Sonnhammer *et al., Proteins 28*:405-420 (1997)) can be used.

PROSITE and BLOCKS represent collected families of protein motifs. Thus, searching these databases entails submitting a single sequence to determine whether or not that sequence is similar to the members of an established family. Programs working in the opposite direction compare a collection of sequences with individual entries in the protein databases. An example of such a program is the Motif Search Tool, or MoST (Tatusov *et al., Proc. Natl. Acad. Sci. (U.S.A.) 91:*12091-12095 (1994), the entirety of which is herein incorporated by reference). On the basis of an aligned set of input sequences, a weight matrix is calculated by using one of four methods (selected by the user); a weight matrix is simply a representation, position by position in an alignment, of how likely a particular amino acid will appear. The calculated weight matrix is then used to search the databases. To increase sensitivity, newly found sequences are added to the original data set, the weight matrix is recalculated, and the search is performed again. This procedure continues until no new sequences are found.

III. Contig Assembly

A characteristic feature of a large scale shotgun sequencing project is that the sequence data can be processed and assembled into contiguous sequences (contigs), which represent a reconstruction of the original genome sequence from the cloned fragments. Programs are available in the public domain that can analyze the sequence output and assemble the sequences into larger sequence regions representing contiguous sequences of the target genome. Examples of such programs can be found at, for example, http://genome.wustl.edu/gsc, http://www.sanger.ac.uk, and http://www.mbt.washington.edu. An example of sequence reading program is Phred (http://www.mbt.washington.edu). Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files.

The process of assembling DNA sequence fragments generally involves three phases; the overlap phase, the layout phase and the multi-alignment, or consensus, phase. In the overlap phase, each fragment is compared against every other fragment to determine if they share a common subsequence, an indication that they were potentially sampled from overlapping stretches of the original DNA strand. Pairs of fragments are compared in two ways; 1) with both fragments in the same relative orientation, and 2) with one of the fragments having been reverse complemented. In the layout phase, a series of alternate assemblies or layouts of the fragments based on the pairwise overlaps is generated. A layout specifies the relative locations and orientations of the fragments with respect to each other and is typically visualized as an arrangement of overlapping directed lines, one for each fragment. The general criterion for the layout phase is to produce plausible assemblies of maximum likelihood. In this manner, it can be determined if there is more than one way to put the pieces together and if different solutions appear equally plausible. In such a case, one would return to the lab and obtain additional information to resolve the ambiguity. The multi-alignment, or consensus, phase uses

more information than just the pairwise alignments in the layout. The sequences of all the fragments in a layout are simultaneously aligned, giving a final set of contigs representing regions of the target genome. An example of an assembly program is PHRAP, which can be found at http://chimera.biotech.washington.edu/UWGC/tools/phrap.htm.

IV. Gene Mapping and Marker Assisted Introgression of Plant Traits

Genome sequence information from *Arabidopsis* provides markers that will assist in the development of improved plants. Marker assisted introgression of traits into plants have been reported. An initial step in that process is the localization of the trait by gene mapping. Gene mapping is the process of determining a gene's position relative to other genes and genetic markers through linkage analysis. The basic principle for linkage mapping is that the closer together two genes are on the chromosome, the more likely they are to be inherited together (Rothwell, *Understanding Genetics. 4th Ed.* Oxford University Press, New York, p. 703 (1988), the entirety of which is herein incorporated by reference). Briefly, a cross is made between two genetically compatible but divergent parents relative to traits under study. Genetic markers are then used to follow the segregation of traits under study in the progeny from the cross (often a backcross, F_2, or recombinant inbred population).

Linkage analysis is based on the level at which markers and genes are co-inherited (Rothwell, *Understanding Genetics. 4th Ed.* Oxford University Press, New York, p. 703 (1988)). Statistical tests like chi-square analysis can be used to test the randomness of segregation or linkage (Kochert, *The Rockefeller Foundation International Program on Rice Biotechnology,* University of Georgia Athens, GA, pp. 1-14 (1989), the entirety of which is herein incorporated by reference). In linkage mapping, the proportion of recombinant individuals out of the total mapping population provides the information for determining the genetic distance between the loci (Young, *Encyclopedia of Agricultural*

Science, Vol. 3, pp. 275-282 (1994), the entirety of which is herein incorporated by reference).

Classical mapping studies utilize easily observable, visible traits instead of molecular markers. These visible traits are also known as naked eye polymorphisms. These traits can be morphological like plant height, fruit size, shape and color or physiological like disease response, photoperiod sensitivity or crop maturity. Visible traits are useful and are still in use because they represent actual phenotypes and are easy to score without any specialized lab equipment. By contrast, the other types of genetic markers are arbitrary loci for use in linkage mapping and often not associated to specific plant phenotypes (Young, *Encyclopedia of Agricultural Science,* Vol. 3, pp. 275-282 (1994). Many morphological markers cause such large effects on phenotype that they are undesirable in breeding programs. Many other visible traits have the disadvantage of being developmentally regulated (*i.e.*, expressed only certain stages; or at specific tissue and organs). Oftentimes, visible traits mask the effects of linked minor genes making it nearly impossible to identify desirable linkages for selection (Tanksely *et al., Biotech.* 7:257-264 (1989), the entirety of which is herein incorporated by reference).

Although a number of important agronomic characters are controlled by loci having major effects on phenotype, many economically important traits, such as yield and some forms of disease resistance, are quantitative in nature. This type of phenotypic variation in a trait is typically characterized by continuous, normal distribution of phenotypic values in a particular population (polygenic traits) (Beckmann and Soller, *Oxford Surveys of Plant Molecular Biology, Miffen.* (ed.), Vol. 3, Oxford University Press, UK., pp. 196-250 (1986), the entirety of which is herein incorporated by reference). Loci contributing to such genetic variation are often termed, minor genes, as opposed to major genes with large effects that follow a Mendelian pattern of inheritance. Polygenic traits are also predicted to follow a Mendelian type of inheritance, however the contribution of each locus is expressed as an increase or decrease in the final trait value.

The advent of DNA markers, such as restriction fragment length polymorphic markers (RFLPs), microsatellite markers, single nucleotide polymorphic markers (SNPs), and random amplified polymorphic markers (RAPDs), allow the resolution of complex, multigenic traits into their individual Mendelian components (Paterson *et al.*, *Nature 335:*721-726 (1988), the entirety of which is herein incorporated by reference). A number of applications of RFLPs and other markers have been suggested for plant breeding. Among the potential applications for RFLPs and other markers in plant breeding include: varietal identification (Soller and Beckmann, *Theor. Appl. Genet. 67:*25-33 (1983), the entirety of which is herein incorporated by reference; Tanksley *et al.*, *Biotech. 7:*257-264 (1989), QTL mapping (Edwards *et al.*, *Genetics 116:*113-115 (1987), the entirety of which is herein incorporated by reference); Nienhuis *et al.*, *Crop Sci. 27:*797-803 (1987); Osborn *et al.*, *Theor. Appl. Genet. 73:*350-356 (1987); Romero-Severson *et al.*, *Use of RFLPs In Analysis Of Quantitative Trait Loci In Maize, In* Helentjaris and Burr (eds.), pp. 97-102 (1989), the entirety of which is herein incorporated by reference; Young *et al.*, *Genetics 120:*579-585 (1988), the entirety of which is herein incorporated by reference; Martin *et al.*, *Science 243:*1725-1728 (1989), the entirety of which is herein incorporated by reference; Sarfatti *et al.*, *Theor. Appl. Genet. 78:*22-26 (1989), the entirety of which is herein incorporated by reference; Tanksley *et al.*, *Biotech. 7:*257-264 (1989); Barone *et al.*, *Mol. Gen. Genet. 224:*177-182 (1990), the entirety of which is herein incorporated by reference; Jung *et al.*, *Theor. Appl. Genet. 79:*663-672 (1990), the entirety of which is herein incorporated by reference; Keim *et al.*, *Genetics 126:*735-742 (1990), the entirety of which is herein incorporated by reference; Keim *et al.*, *Theor. Appl. Genet. 79:*465-369 (1990), the entirety of which is herein incorporated by reference; *Paterson et al.*, *Genetics 124:*735-742 (1990), the entirety of which is herein incorporated by reference; Martin *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 88:*2336-2340 (1991), the entirety of which is herein incorporated by reference; Messeguer *et al.*, *Theor. Appl. Genet. 82:*529-536 (1991), the entirety of which is herein

incorporated by reference; Michelmore *et al., Proc. Natl. Acad. Sci. (U.S.A.) 88:*9828-9832 (1991), the entirety of which is herein incorporated by reference; Ottaviano *et al., Theor. Appl. Genet. 81:*713-719 (1991), the entirety of which is herein incorporated by reference; Yu *et al., Theor. Appl. Genet. 81:*471-476 (1991), the entirety of which is herein incorporated by reference; Diers *et al., Crop Sci. 32:*377-383 (1992), the entirety of which is herein incorporated by reference; *Theor. Appl. Genet. 83:*608-612 (1992), the entirety of which is herein incorporated by reference; *J. Plant Nut. 15:*2127-2136 (1992), the entirety of which is herein incorporated by reference; Doebley *et al., Proc. Natl. Acad. Sci. (U.S.A.) 87:*9888-9892 (1990), the entirety of which is herein incorporated by reference, screening genetic resource strains for useful quantitative trait alleles and introgression of these alleles into commercial varieties (Beckmann and Soller, *Theor. Appl. Genet. 67:*35-43 (1983), the entirety of which is herein incorporated by reference; marker-assisted selection (Tanksley *et al., Biotech. 7:*257-264 (1989) and map-based cloning (Tanksley *et al., Biotech. 7:*257-264 (1989). In addition, DNA markers can be used to obtain information about: (1) the number, effect, and chromosomal location of each gene affecting a trait; (2) effects of multiple copies of individual genes (gene dosage); (3) interaction between/among genes controlling a trait (epistasis); (4) whether individual genes affect more than one trait (pleiotropy); and (5) stability of gene function across environments (G x E interactions).

Summary of the Invention

The present invention provides a substantially purified nucleic acid molecule, the nucleic acid molecule capable of specifically hybridizing to a second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either.

The present invention also provides a substantially purified nucleic acid molecule encoding an *Arabidopsis* protein or fragment thereof, wherein the *Arabidopsis* protein or

fragment thereof is encoded by a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through ID NO: 27,741 or complements thereof or fragments of either.

The present invention also provides a substantially purified protein or fragment thereof encoded by a first nucleic acid molecule which specifically hybridizes to a second nucleic acid molecule, the second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof.

The present invention also provides a substantially purified protein or fragment thereof encoded by a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof or fragments of either.

The present invention also provides a substantially purified antibody or fragment thereof, the antibody or fragment thereof capable of specifically binding to the protein or fragment thereof encoded by a first nucleic acid molecule which specifically hybridizes to a second nucleic acid molecule, the second nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof or fragments of either.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of an mRNA molecule; which is linked to (B) a structural nucleic acid molecule, wherein the structural nucleic acid molecule is selected from the group consisting of a protein or fragment thereof encoding sequence located within SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof; which is linked to (C) a 3' non-translated sequence that functions in a plant cell to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant

cell to cause the production of an mRNA molecule wherein the promoter nucleic acid molecule is selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either; which is linked to (B) a structural nucleic acid molecule encoding a protein or fragment thereof; which is linked to (C) a 3' non-translated sequence that functions in a plant cell to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of an mRNA molecule; which is linked to (B) a transcribed nucleic acid molecule with a transcribed strand and a non-transcribed strand, wherein the transcribed strand is complementary to a nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof and the transcribed strand is complementary to an endogenous mRNA molecule; which is linked to (C) a 3' non-translated sequence that functions in plant cells to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

The present invention also provides a transformed plant having a nucleic acid molecule which comprises: (A) an exogenous promoter region which functions in a plant cell to cause the production of an mRNA molecule wherein the promoter nucleic acid molecule is selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof; which is linked to (B) a transcribed nucleic acid molecule with a transcribed strand and a non-transcribed strand, wherein the transcribed strand is complementary to an endogenous mRNA molecule; which is linked to (C) a 3' non-translated sequence that functions in plant cells to cause termination of transcription and addition of polyadenylated ribonucleotides to a 3' end of the mRNA molecule.

The present invention also provides a computer readable medium having recorded thereon one or more of the nucleotide sequences depicted in SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof.

The present invention also provides a computer readable medium having recorded thereon one or more nucleic acid molecules encoding an *Arabidopsis* protein or fragment thereof, wherein the *Arabidopsis* protein or fragment thereof is encoded by a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either.

The present invention also provides a method of introgressing a trait into a plant comprising using a nucleic acid marker for marker assisted selection of the plant, the nucleic acid marker complementary to a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof, and introgressing the trait into a plant.

The present invention also provides a method for screening for a trait comprising interrogating genomic DNA for the presence or absence of a marker molecule that is genetically linked to a nucleic acid sequence complementary to a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof; and detecting the presence or absence of the marker.

The present invention also provides a method for determining the likelihood of the presence or absence of a trait in a plant comprising the steps of: (A) obtaining genomic DNA from the plant; (B) detecting a marker nucleic acid molecule; wherein the marker nucleic acid molecule specifically hybridizes with a nucleic acid sequence that is genetically linked to a nucleic acid sequence complementary to a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof; (C) determining the level, presence or absence of the marker nucleic acid molecule, wherein the level, presence or absence of the marker nucleic acid molecule is indicative of the likely presence in the plant of the trait.

The present invention also provides a method for determining a genomic polymorphism in a plant that is predictive of a trait comprising the steps: (A) incubating a marker nucleic acid molecule, under conditions permitting nucleic acid hybridization, and a complementary nucleic acid molecule obtained from the plant, the marker nucleic acid molecule having a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either; (B) permitting hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule obtained from the plant; and (C) detecting the presence of the polymorphism.

The present invention also provides a method of determining an association between a polymorphism and a plant trait comprising: (A) hybridizing a nucleic acid molecule specific for the polymorphism to genetic material of a plant, wherein the nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either; and (B) calculating the degree of association between the polymorphism and the plant trait.

Detailed Description of the Invention

Agents of the invention:

(a) Nucleic Acid Molecules

Agents of the present invention include plant nucleic acid molecules and more specifically include *Arabidopsis* nucleic acid molecules, and more specifically include *Arabidopsis thaliana, Columbia* nucleic acid molecules. A subset of the nucleic acid molecules of the present invention includes nucleic acid molecules that are marker molecules. Another subset of the nucleic molecules of the present invention includes nucleic acid molecules that are promoters and/or regulatory elements. Another subset of the nucleic acid molecules of the present invention includes nucleic acid molecules that encode a gene or fragment thereof. Another subset of the nucleic acid molecules of the

present invention encodes proteins or fragments of proteins. In a preferred embodiment the nucleic acid molecules of the present invention are derived from *Arabidopsis* and in an even more preferred embodiment the nucleic acid molecules of the present invention are derived from *Arabidopsis thaliana, Columbia.*

Fragment nucleic acid molecules may encode significant portion(s) of, or indeed most of, these nucleic acid molecules. For example, a fragment nucleic acid molecule can encode an *Arabidopsis* protein or fragment thereof. Alternatively, the fragments may comprise smaller oligonucleotides (having from about 15 to about 250 nucleotide residues, and more preferably, about 15 to about 30 nucleotide residues).

As used herein, an agent, be it a naturally occurring molecule or otherwise may be "substantially purified," if desired, referring to a molecule separated from substantially all other molecules normally associated with it in its native state. More preferably a substantially purified molecule is the predominant species present in a preparation. A substantially purified molecule may be greater than 60% free, preferably 75% free, more preferably 90% free, and most preferably 95% free from the other molecules (exclusive of solvent) present in the natural mixture. The term "substantially purified" is not intended to encompass molecules present in their native state.

The agents of the present invention will preferably be "biologically active" with respect to either a structural attribute, such as the capacity of a nucleic acid to hybridize to another nucleic acid molecule, or the ability of a protein to be bound by an antibody (or to compete with another molecule for such binding). Alternatively, such an attribute may be catalytic, and thus involve the capacity of the agent to mediate a chemical reaction or response.

The agents of the present invention may also be recombinant. As used herein, the term recombinant means any agent (*e.g.*, DNA, peptide *etc.*), that is, or results, however indirect, from human manipulation of a nucleic acid molecule.

It is understood that the agents of the present invention may be labeled with reagents that facilitate detection of the agent (*e.g.*, fluorescent labels, Prober *et al.*, *Science 238*:336-340 (1987); Albarella *et al.*, EP 144914, chemical labels, Sheldon *et al.*, U.S. Patent 4,582,789; Albarella *et al.*, U.S. Patent 4,563,417, modified bases, Miyoshi *et al.*, EP 119448, all of which are hereby incorporated by reference in their entirety).

It is further understood, that the present invention provides recombinant bacterial, viral, microbial, insect, fungal and plant cells comprising the agents of the present invention.

Nucleic acid molecules or fragments thereof of the present invention are capable of specifically hybridizing to other nucleic acid molecules under certain circumstances. As used herein, two nucleic acid molecules are said to be capable of specifically hybridizing to one another if the two molecules are capable of forming an anti-parallel, double-stranded nucleic acid structure. A nucleic acid molecule is said to be the "complement" of another nucleic acid molecule if they exhibit complete complementarity. As used herein, molecules are said to exhibit "complete complementarity" when every nucleotide of one of the molecules is complementary to a nucleotide of the other. Two molecules are said to be "minimally complementary" if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under at least conventional "low-stringency" conditions. Similarly, the molecules are said to be "complementary" if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under conventional "high-stringency" conditions. Conventional stringency conditions are described by Sambrook *et al.*, *Molecular Cloning,* A Laboratory Manual, 2nd Ed., Cold Spring Harbor Press, Cold Spring Harbor, New York (1989), and by Haymes *et al. Nucleic Acid Hybridization, A Practical Approach*, IRL Press, Washington, DC (1985), the entirety of which is herein incorporated by reference. Departures from complete complementarity are therefore permissible, as long as such departures do not completely preclude the

capacity of the molecules to form a double-stranded structure. Thus, in order for a nucleic acid molecule to serve as a primer or probe it need only be sufficiently complementary in sequence to be able to form a stable double-stranded structure under the particular solvent and salt concentrations employed.

Appropriate stringency conditions which promote DNA hybridization, for example, 6.0 X sodium chloride/sodium citrate (SSC) at about 45°C, followed by a wash of 2.0 X SSC at 50°C, are known to those skilled in the art or can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6. For example, the salt concentration in the wash step can be selected from a low stringency of about 2.0 X SSC at 50°C to a high stringency of about 0.2 X SSC at 50°C. In addition, the temperature in the wash step can be increased from low stringency conditions at room temperature, about 22°C, to high stringency conditions at about 65°C. Both temperature and salt may be varied, or either the temperature or the salt concentration may be held constant while the other variable is changed.

In a preferred embodiment, a nucleic acid of the present invention will specifically hybridize to one or more of the nucleic acid molecules set forth in SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof under moderately stringent conditions, for example at about 2.0 X SSC and about 65°C.

In a particularly preferred embodiment, a nucleic acid of the present invention will include those nucleic acid molecules that specifically hybridize to one or more of the nucleic acid molecules set forth in SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof under high stringency conditions.

In one aspect of the present invention, the nucleic acid molecules of the present invention have one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof. In another aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 90% sequence identity with one or more of the nucleic acid sequences

set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof. In a further aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 95% sequence identity with one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof. In a more preferred aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 98% sequence identity with one or more of the nucleic acid sequences set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof. In an even more preferred aspect of the present invention, one or more of the nucleic acid molecules of the present invention share between 100% and 99% sequence identity with one or more of the sequences set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof.

(i) Nucleic Acid Molecule Markers

One aspect of the present invention concerns nucleic acid molecules SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof or fragments of either, that contain microsatellites or other markers. Microsatellites, also known as SSRs, typically include 1-6 nucleotide core element within SEQ ID NO: 1 through SEQ ID NO: 51,470 that are tandemly repeated from one to many thousands of times. A different "allele" occurs at an SSR locus as a result of changes in the number of times a core element is repeated, altering the length of the repeat region, (Brown *et al., Methods of Genome Analysis in Plants*, (ed.) Jauhar, CRC Press, Inc, Boca Raton, Florida, USA; London, England, UK, pp. 147-159, (1996), the entirety of which is herein incorporated by reference).

SSR loci occur throughout plant genomes, and specific repeat motifs occur at different levels of abundance than those found in animals. The relative frequencies of all SSRs with repeat units of 1-6 nucleotides have been surveyed. The most abundant SSR is AAAAAT followed by A_n, AG_n AAT, AAC, AGC, AAG, AATT, AAAT and AC. On

average, 1 SSR is found every 21 and 65 kb in dicots and monocots. Fewer CG nucleotides are found in dicots than in monocots. There is no reported correlation between abundance of SSRs and nuclear DNA content. The abundance of all tri and tetranucleotide SSR combinations jointly have been reported to be equivalent to that of the total di-nucleotide combinations. Mono- di- and tetra-nucleotide repeats are all located in noncoding regions of DNA while 57% of those trinucleotide SSRs containing CG were located within gene coding regions. All repeated trinucleotide SSRs composed entirely of AT are found in noncoding regions, (Brown *et al., Methods of Genome Analysis in Plants, ed.* Jauhar, CRC Press, Inc, Boca Raton, Florida, USA; London, England, UK, pp. 147-159, (1996).

Microsatellites can be observed in SEQ NO:1 to SEQ NO: 51,470 or complements thereof by using the BLASTN program to examine sequences for the presence/absence of microsatellites. In this system, raw sequence data is searched through SSR databases, SSRDB1 and SSRDB2 using the BLAST program. SSRDB1 stores SSR markers collected from publications while the public database contains 692 classes of di-, tri and tetranucleotide repeat markers generated by computer.

Microsatellites can also be observed by screening the nucleic acid molecules of the present invention by colony or plaque hybridization with a labeled probe containing microsatellite markers; isolating positive clones and sequencing the inserts of the positive clones; suitable primers flanking the microsatellite markers.

Genetic markers of the present invention include "dominant" or "codominant" markers. "Codominant markers" reveal the presence of two or more alleles (two per diploid individual) at a locus. "Dominant markers" reveal the presence of only a single allele per locus. The presence of the dominant marker phenotype (*e.g.*, a band of DNA) is an indication that one allele is present in either the homozygous or heterozygous condition. The absence of the dominant marker phenotype (*e.g.*, absence of a DNA band) is merely evidence that "some other" undefined allele is present. In the case of

populations where individuals are predominantly homozygous and loci are predominately dimorphic, dominant and codominant markers can be equally valuable. As populations become more heterozygous and muti-allelic, codominant markers often become more informative of the genotype than dominant markers.

Additional markers, such as AFLP markers, RFLP markers, RAPD markers, SNPs, phenotypic markers, isozyme markers can be utilized (Walton, Seed World 22-29 (July, 1993), the entirety of which is herein incorporated by reference; Burow and Blake, *Molecular Dissection of Complex Traits*, 13-29, Eds. Paterson, CRC Press, New York (1988), the entirety of which is herein incorporated by reference). DNA markers can be developed from nucleic acid molecules using restriction endonucleases, the PCR and/or DNA sequence information. RFLP markers result from single base changes or insertions/deletions. These codominant markers are highly abundant in plant genomes, have a medium level of polymorphism and are developed by a combination of restriction endonuclease digestion and Southern blotting hybridization. CAPS are similarly developed from restriction nuclease digestion but only of specific PCR products. These markers are also codominant, have a medium level of polymorphism and are highly abundant in the genome. The CAPS result from single base changes and insertions/deletions. Another marker type, RAPDs, are developed from DNA amplification with random primers and result from single base changes and insertions/deletions in plant genomes. They are dominant markers with a medium level of polymorphisms and are highly abundant. AFLP markers require using the PCR on a subset of restriction fragments from extended adapter primers. These markers are both dominant and codominant are highly abundant in genomes and exhibit a medium level of polymorphism. SSRs require DNA sequence information. These codominant markers result from repeat length changes, are highly polymorphic, and do not exhibit as high a degree of abundance in the genome as CAPS, AFLPs and RAPDs. SNPs also require DNA sequence information. These codominant markers result from single base

substitutions. They are highly abundant and exhibit a medium of polymorphism (Rafalski *et al.*, In: *Nonmammalian Genomic Analysis*, ed. Birren and Lai, Academic Press, San Diego, CA, pp. 75-134 (1996), the entirety of which is herein incorporated by reference). Methods to isolate such markers are known in the art.

(ii) Nucleic Acid Molecules Comprising Regulatory Elements

Another class of agents of the present invention are nucleic acid molecules having promoter regions or partial promoter regions or regulatory elements, particularly those promoter regions or partial promoter regions or regulatory elements located within SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements thereof. Such promoter regions are typically found upstream of the trinucleotide ATG sequence at the start site of a protein coding region.

As used herein, a promoter region is a region of a nucleic acid molecule that is capable, when located in *cis* to a nucleic acid sequence that encodes for a protein or peptide to function in a way that directs expression of one or more mRNAs molecules that encodes for the protein or peptide.

Promoters of the present invention can include between about 300bp upstream and about 10kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region. Promoters of the present invention can preferably include between about 300bp upstream and about 5kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region. Promoters of the present invention can more preferably include between about 300bp upstream and about 2kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region. Promoters of the present invention can include between about 300bp upstream and about 1kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region. While in many circumstances a 300bp promoter may be sufficient for expression, additional sequences may act to further regulate expression, for example, in response to biochemical, developmental or environmental signals.

It is also preferred that the promoters of the present invention contain a CAAT and a TATA cis element. Moreover, the promoters of the present invention can contain one or more *cis* elements in addition to a CAAT and a TATA box.

By "regulatory element" it is intended a series of nucleotides that determines if, when, and at what level a particular gene is expressed. The regulatory DNA sequences specifically interact with regulatory or other proteins. Many regulatory elements act in *cis* ("*cis* elements") and are believed to affect DNA topology, producing local conformations that selectively allow or restrict access of RNA polymerase to the DNA template or that facilitate selective opening of the double helix at the site of transcriptional initiation. *Cis* elements occur within, but are not limited to promoters, and promoter modulating sequences (inducible elements). *Cis* elements can be identified using known *cis* elements as a target sequence or target motif in the BLAST programs of the present invention.

Promoters of the present invention include homologues of *cis* elements known to effect gene regulation that show homology with the nucleic acid molecules of the present invention. These *cis* elements include, but are not limited to, oxygen responsive *cis* elements (Cowen *et al.*, *J Biol. Chem. 268:*26904-26910 (1993) the entirety of which is herein incorporated by reference), light regulatory elements (Bruce and Quaill, *Plant Cell 2:*1081-1089 (1990) the entirety of which is herein incorporated by reference; Bruce *et al., EMBO J. 10:*3015-3024 (1991), the entirety of which is herein incorporated by reference; Rocholl *et al., Plant Sci. 97:*189-198 (1994), the entirety of which is herein incorporated by reference; Block *et al., Proc. Natl. Acad. Sci. (U.S.A.) 87:*5387-5391 (1990), the entirety of which is herein incorporated by reference; Giuliano *et al., Proc. Natl. Acad. Sci. (U.S.A.) 85:*7089-7093 (1988), the entirety of which is herein incorporated by reference; Staiger *et al., Proc. Natl. Acad. Sci. (U.S.A.) 86:*6930-6934 (1989), the entirety of which is herein incorporated by reference; Izawa *et al., Plant Cell 6:*1277-1287 (1994), the entirety of which is herein incorporated by reference; Menkens *et al., Trends in Biochemistry 20:*506-510 (1995), the entirety of which is herein

incorporated by reference; Foster *et al., FASEB J. 8:*192-200 (1994), the entirety of which is herein incorporated by reference; Plesse *et al., Mol Gen Gene 254:*258-266 (1997), the entirety of which is herein incorporated by reference; Green *et al., EMBO J. 6:*2543-2549 (1987), the entirety of which is herein incorporated by reference; Kuhlemeier *et al., Ann. Rev Plant Physiol. 38:*221-257 (1987), the entirety of which is herein incorporated by reference; Villain *et al., J. Biol. Chem. 271:*32593-32598 (1996), the entirety of which is herein incorporated by reference; Lam *et al., Plant Cell 2:*857-866 (1990), the entirety of which is herein incorporated by reference; Gilmartin *et al., Plant Cell 2:*369-378 (1990), the entirety of which is herein incorporated by reference; Datta *et al., Plant Cell 1:*1069-1077 (1989) the entirety of which is herein incorporated by reference; Gilmartin *et al., Plant Cell 2:*369-378 (1990), the entirety of which is herein incorporated by reference; Castresana *et al., EMBO J.* 7:1929-1936 (1988), the entirety of which is herein incorporated by reference; Ueda *et al., Plant Cell 1:*217-227 (1989), the entirety of which is herein incorporated by reference; Terzaghi *et al., Annu. Rev. Plant Physiol. Plant Mol. Biol. 46:*445-474 (1995), the entirety of which is herein incorporated by reference; Green *et al., EMBO J.* 6:2543-2549 (1987), the entirety of which is herein incorporated by reference; Villain *et al., J. Biol. Chem. 271:*32593-32598 (1996), the entirety of which is herein incorporated by reference; Tjaden *et al., Plant Cell 6:*107-118 (1994), the entirety of which is herein incorporated by reference; Tjaden *et al., Plant Physiol. 108:*1109-1117 (1995), the entirety of which is herein incorporated by reference; Ngai *et al., Plant J. 12:*1021-1234 (1997), the entirety of which is herein incorporated by reference; Bruce *et al., EMBO J. 10:*3015-3024 (1991), the entirety of which is herein incorporated by reference; Ngai *et al., Plant J. 12:*1021-1034 (1997), the entirety of which is herein incorporated by reference), elements responsive to gibberellin, (Muller *et al., J. Plant Physiol. 145:*606-613 (1995), the entirety of which is herein incorporated by reference; Croissant *et al., Plant Science 116:*27-35 (1996), the entirety of which is herein incorporated by reference; Lohmer *et al., EMBO J. 10:*617-624 (1991), the entirety of

which is herein incorporated by reference; Rogers *et al., Plant Cell 4:*1443-1451 (1992), the entirety of which is herein incorporated by reference; Lanahan *et al., Plant Cell* 4:203-211 (1992) the entirety of which is herein incorporated by reference; Skriver *et al., Proc. Natl. Acad. Sci. (U.S.A.) 88:*7266-7270 (1991) the entirety of which is herein incorporated by reference; Gilmartin *et al., Plant Cell 2:*369-378 (1990), the entirety of which is herein incorporated by reference; Huang *et al., Plant Mol. Biol. 14:*655-668 (1990), the entirety of which is herein incorporated by reference, Gubler *et al., Plant Cell 7:*1879-1891 (1995), the entirety of which is herein incorporated by reference), elements responsive to abscisic acid, (Busk *et al., Plant Cell 9:*2261-2270 (1997), the entirety of which is herein incorporated by reference; Guiltinan *et al., Science 250:*267-270 (1990), the entirety of which is herein incorporated by reference; Shen *et al., Plant Cell 7:*295-307 (1995) the entirety of which is herein incorporated by reference; Shen *et al., Plant Cell 8:*1107-1119 (1996), the entirety of which is herein incorporated by reference; Seo *et al., Plant Mol. Biol. 27:*1119-1131 (1995), the entirety of which is herein incorporated by reference; Marcotte *et al., Plant Cell* 1:969-976 (1989) the entirety of which is herein incorporated by reference; Shen *et al., Plant Cell* 7:295-307 (1995), the entirety of which is herein incorporated by reference; Iwasaki *et al., Mol Gen Genet* 247:391-398 (1995), the entirety of which is herein incorporated by reference; Hattori *et al., Genes Dev.* 6:609-618 (1992), the entirety of which is herein incorporated by reference; Thomas *et al., Plant Cell* 5:1401-1410 (1993), the entirety of which is herein incorporated by reference), elements similar to abscisic acid responsive elements, (Ellerstrom *et al., Plant Mol. Biol. 32:*1019-1027 (1996), the entirety of which is herein incorporated by reference), auxin responsive elements (Liu *et al., Plant Cell 6:*645-657 (1994) the entirety of which is herein incorporated by reference; Liu *et al., Plant Physiol.* 115:397-407 (1997), the entirety of which is herein incorporated by reference; Kosugi *et al., Plant J.* 7:877-886 (1995), the entirety of which is herein incorporated by reference; Kosugi *et al., Plant Cell* 9:1607-1619 (1997), the entirety of which is herein incorporated by reference; Ballas *et*

*al., J. Mol. Biol. 233:*580-596 (1993), the entirety of which is herein incorporated by reference), a *cis* element responsive to methyl jasmonate treatment (Beaudoin and Rothstein, *Plant Mol. Biol. 33:*835-846 (1997), the entirety of which is herein incorporated by reference), a *cis* element responsive to abscisic acid and stress response (Straub *et al., Plant Mol. Biol.* 26:617-630 (1994), the entirety of which is herein incorporated by reference), ethylene responsive *cis* elements (Itzhaki *et al., Proc. Natl. Acad. Sci. (U.S.A.)* 91:8925-8929 (1994), the entirety of which is herein incorporated by reference; Montgomery *et al., Proc. Natl. Acad. Sci. (U.S.A.) 90:*5939-5943 (1993), the entirety of which is herein incorporated by reference; Sessa *et al., Plant Mol. Biol.* 28:145-153 (1995), the entirety of which is herein incorporated by reference; Shinshi *et al., Plant Mol. Biol.* 27:923-932 (1995), the entirety of which is herein incorporated by reference), salicylic acid *cis* responsive elements, (Strange *et al., Plant J.* 11:1315-1324 (1997), the entirety of which is herein incorporated by reference; Qin *et al., Plant Cell* 6:863-874 (1994), the entirety of which is herein incorporated by reference), a *cis* element that responds to water stress and abscisic acid (Lam *et al., J. Biol. Chem.* 266:17131-17135 (1991), the entirety of which is herein incorporated by reference; Thomas *et al., Plant Cell* 5:1401-1410 (1993), the entirety of which is herein incorporated by reference; Pla *et al., Plant Mol Biol* 21:259-266 (1993), the entirety of which is herein incorporated by reference), a *cis* element essential for M phase-specific expression (Ito *et al., Plant Cell* 10:331-341 (1998), the entirety of which is herein incorporated by reference), sucrose responsive elements (Huang *et al., Plant Mol. Biol.* 14:655-668 (1990), the entirety of which is herein incorporated by reference; Hwang *et al., Plant Mol. Biol.* 36:331-341 (1998), the entirety of which is herein incorporated by reference; Grierson *et al., Plant J.* 5:815-826 (1994), the entirety of which is herein incorporated by reference), heat shock response elements (Pelham *et al., Trends Genet.* 1:31-35 (1985), the entirety of which is herein incorporated by reference), elements responsive to auxin and/or salicylic acid and also reported for light regulation (Lam *et al., Proc. Natl. Acad. Sci.*

(U.S.A.) 86:7890-7897 (1989), the entirety of which is herein incorporated by reference; Benfey *et al., Science* 250:959-966 (1990), the entirety of which is herein incorporated by reference), elements responsive to ethylene and salicylic acid (Ohme-Takagi *et al., Plant Mol. Biol.* 15:941-946 (1990), the entirety of which is herein incorporated by reference), elements responsive to wounding and abiotic stress (Loake *et al., Proc. Natl. Acad. Sci. (U.S.A.)* 89:9230-9234 (1992), the entirety of which is herein incorporated by reference; Mhiri *et al., Plant Mol. Biol.* 33:257-266 (1997), the entirety of which is herein incorporated by reference), antoxidant response elements (Rushmore *et al., J. Biol. Chem.* 266:11632-11639, the entirety of which is herein incorporated by reference; Dalton *et al., Nucleic Acids Res.* 22:5016-5023 (1994), the entirety of which is herein incorporated by reference), Sph elements (Suzuki *et al., Plant Cell 9:*799-807 (1997), the entirety of which is herein incorporated reference), Elicitor responsive elements, (Fukuda *et al., Plant Mol. Biol.* 34:81-87 (1997), the entirety of which is herein incorporated by reference; Rushton *et al., EMBO J.* 15:5690-5700 (1996), the entirety of which is herein incorporated by reference), metal responsive elements (Stuart *et al., Nature* 317:828-831 (1985), the entirety of which is herein incorporated by reference; Westin *et al., EMBO J.* 7:3763-3770 (1988), the entirety of which is herein incorporated by reference; Thiele *et al., Nucleic Acids Res.* 20:1183-1191 (1992), the entirety of which is herein incorporated by reference; Faisst *et al., Nucleic Acids Res.* 20:3-26 (1992), the entirety of which is herein incorporated by reference), low temperature responsive elements, (Baker *et al., Plant Mol. Biol.* 24:701-713 (1994), the entirety of which is herein incorporated by reference; Jiang *et al., Plant Mol. Biol.* 30:679-684 (1996), the entirety of which is herein incorporated by reference; Nordin *et al., Plant Mol. Biol.* 21:641-653 (1993), the entirety of which is herein incorporated by reference; Zhou *et al., J. Biol. Chem.* 267:23515-23519 (1992), the entirety of which is herein incorporated by reference), drought responsive elements, (Yamaguchi *et al., Plant Cell* 6:251-264 (1994), the entirety of which is herein incorporated by reference; Wang *et al., Plant Mol. Biol.* 28:605-617

(1995), the entirety of which is herein incorporated by reference; Bray, *Trends in Plant Science* 2:48-54 (1997), the entirety of which is herein incorporated by reference), enhancer elements for glutenin, (Colot *et al., EMBO J.* 6:3559-3564 (1987), the entirety of which is herein incorporated by reference; Thomas *et al., Plant Cell* 2:1171-1180 (1990), the entirety of which is incorporated by reference; Kreis *et al., Philos. Trans. R. Soc. Lond.,* B314:355-365 (1986), the entirety of which is herein incorporated by reference), light-independent regulatory elements, (Lagrange *et al., Plant Cell* 9:1469-1479 (1997), the entirety of which is herein incorporated by reference; Villain *et al., J. Biol. Chem.* 271:32593-32598 (1996), the entirety of which is herein incorporated by reference), OCS enhancer elements, (Bouchez *et al., EMBO J.* 8:4197-4204 (1989), the entirety of which is herein incorporated by reference; Foley *et al., Plant J.* 3:669-679 (1993), the entirety of which is herein incorporated by reference), ACGT elements, (Foster *et al., FASEB J.* 8:192-200 (1994), the entirety of which is herein incorporated by reference; Izawa *et al., Plant Cell* 6:1277-1287 (1994), the entirety of which is herein incorporated by reference; Izawa *et al., J. Mol. Biol.* 230:1131-1144 (1993) the entirety of which is herein incorporated by reference), negative *cis* elements in plastid related genes, (Zhou *et al., J. Biol. Chem.* 267:23515-23519 (1992), the entirety of which is herein incorporated by reference; Lagrange *et al., Mol. Cell Biol.* 13:2614-2622 (1993), the entirety of which is herein incorporated by reference; Lagrange *et al., Plant Cell* 9:1469-1479 (1997), the entirety of which is herein incorporated by reference; Zhou *et al. J. Biol. Chem.* 267:23515-23519 (1992), the entirety of which is herein incorporated by reference), prolamin box elements, (Forde *et al., Nucleic Acids Res.* 13:7327-7339 (1985), the entirety of which is herein incorporated by reference; Colot *et al., EMBO J.* 6:3559-3564 (1987), the entirety of which is herein incorporated by reference; Thomas *et al., Plant Cell* 2:1171-1180 (1990), the entirety of which is herein incorporated by reference; Thompson *et al., Plant Mol. Biol.* 15:755-764 (1990), the entirety of which is herein incorporated by reference; Vicente *et al., Proc. Natl. Acad. Sci. U.S.A.* 94:7685-

7690 (1997), the entirety of which is herein incorporated by reference), elements in enhancers from the IgM heavy chain gene (Gillies *et al., Cell* 33:717-728 (1983), the entirety of which is herein incorporated by reference; Whittier *et al., Nucleic Acids Res.* 15:2515-2535 (1987), the entirety of which is herein incorporated by reference. Additional promoters and regulatory elements are set forth under the section titled "Uses of the Agents of the Invention."

(iii) Nucleic Acid Molecules Comprising Genes or Fragments Thereof

Nucleic acid molecules of the present invention can comprise one or more genes or fragments thereof. Such genes or fragments thereof include homologues of known genes in other organisms or genes or fragments thereof that elicit only limited or no matches with known genes. Such genes include *Arabidopsis* molecules that encode homologues of known proteins.

Genomic sequences can be screened for the presence of protein homologues or protein coding regions utilizing one or a number of different search algorithms have that been developed, one example of which are the suite of programs referred to as BLAST programs. Other examples of suitable programs that can be utilized are known in the art, several of which are described above in the Background and under the section titled "Uses of the Agents of the Invention."

In a preferred embodiment of the present invention, the nucleic acid molecule encodes an *Arabidopsis* protein or fragment thereof that is a homologue of another plant protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes an *Arabidopsis* protein or fragment thereof that is a homologue of a fungal protein. In another preferred embodiment of the invention, the nucleic acid molecule encodes an *Arabidopsis* protein or fragment thereof that is a homologue of mammalian protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes an *Arabidopsis* protein or fragment thereof that is a homologue of a bacterial protein.

In a preferred embodiment of the present invention, the nucleic acid molecule of the present invention encodes an *Arabidopsis* protein or fragment thereof where the *Arabidopsis* protein exhibits a BLAST probability score of greater than 1E-12, preferably a BLAST probability score of between about 1E-30 and about 1E-12, even more preferably a BLAST probability score of greater than 1E-30 with its homologue. In another preferred embodiment of the present invention, the nucleic acid molecule encoding the *Arabidopsis* protein or fragment thereof or fragment thereof exhibits a % identity with its homologue of between about 25%and about 40%, more preferably of between about 40 and about 70%, even more preferably of between about 70% and about 90%, and even more preferably between about 90% and 99%. In another preferred embodiment, of the present invention, the *Arabidopsis* protein homologue or fragments thereof exhibits a % identity with its homologue of 100%.

Nucleic acid molecules of the present invention also include non-*Arabidopsis* homologues. Preferred non-*Arabidopsis* homologues are selected from the group consisting of alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, maize, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus*. Even more preferred non-*Arabidopsis* homologues are selected from the group consisting of *Brassicaceae*.

In a preferred embodiment, nucleic acid molecules having SEQ ID NO: 1 through SEQ ID NO: 51,470 or complements and fragments of either can be utilized to obtain such homologues.

The degeneracy of the genetic code, which allows different nucleic acid sequences to code for the same protein or peptide, is known in the literature. (U.S. Patent No. 4,757,006, the entirety of which is herein incorporated by reference). As used herein a nucleic acid molecule is degenerate of another nucleic acid molecule when the nucleic

acid molecules encode for the same amino acid sequences but comprise different nucleotide sequences. An aspect of the present invention is that the nucleic acid molecules of the present invention include nucleic acid molecules that are degenerate of those set forth in SEQ ID NO: 1 through to SEQ ID NO: 51,470 or complements thereof.

In a further aspect of the present invention, one or more of the nucleic acid molecules of the present invention differ in nucleic acid sequence from those encoding an *Arabidopsis* homologue of fragment thereof in SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof due to the degeneracy in the genetic code in that they encode the same protein but differ in nucleic acid sequence. In another further aspect of the present invention, one or more of the nucleic acid molecules of the present invention differ in nucleic acid sequence from those encoding an *Arabidopsis* homologue of fragment thereof in SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof due to fact that the different nucleic acid sequence encodes a protein having one or more conservative amino acid residue. Examples of conservative substitutions are set forth in Table 1. It is understood that codons capable of coding for such conservative substitutions are known in the art.

Table 1

Original Residue	Conservative Substitutions
Ala	ser
Arg	lys
Asn	gln; his
Asp	glu
Cys	ser; ala
Gln	asn
Glu	asp
Gly	pro
His	asn; gln

Ile	leu; val
Leu	ile; val
Lys	arg; gln; glu
Met	leu; ile
Phe	met; leu; tyr
Ser	thr
Thr	ser
Trp	tyr
Tyr	trp; phe
Val	ile; leu

(iv) Nucleic Acid Molecules Comprising Introns and/or Intron/Exon Junctions

Nucleic acid molecules of the present invention can comprise an intron and/or one or more intron/exon junctions. Sequences of the present invention can be screened for introns and intron/exon junctions utilizing one or a number of different search algorithms have that been developed, one example of which are the suite of programs referred to as BLAST programs. Other examples of suitable programs that can be utilized are known in the art, several of which are described above in the Background and in the section entitled "Uses of the Agents of the Present Invention."

(b) Protein and Peptide Molecules

A class of agents comprises one or more of the protein or peptide molecules encoded by an gene or fragment thereof located within SEQ ID NO: 1 through SEQ ID NO: 27,741, fragments thereof or complements of either or one or more of the protein encoded by a nucleic acid molecule or fragment thereof or peptide molecules encoded by other nucleic acid agents of the present invention. Protein and peptide molecules can be identified using known protein or fragment molecules as a target sequence or target motif

in the BLAST programs of the present invention. In a preferred embodiment the protein or peptide molecules of the present invention are derived from *Arabidopsis* and more preferably *Arabidopsis thaliana, Columbia.*

As used herein, the term "protein molecule" or "peptide molecule" includes any molecule that comprises five or more amino acids. It is well known in the art that proteins or peptides may undergo modification, including post-translational modifications, such as, but not limited to, disulfide bond formation, glycosylation, phosphorylation, or oligomerization. Thus, as used herein, the term "protein molecule" or "peptide molecule" includes any protein molecule that is modified by any biological or non-biological process. The terms "amino acid" and "amino acids" refer to all naturally occurring L-amino acids. This definition is meant to include norleucine, ornithine, homocysteine, and homoserine.

One or more of the protein or fragment of peptide molecules may be produced via chemical synthesis, or more preferably, by expression in a suitable bacterial or eukaryotic host. Suitable methods for expression are described by Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual, 2nd Edition,* Cold Spring Harbor Press, Cold Spring Harbor, New York (1989), or similar texts.

A "protein fragment" is a peptide or polypeptide molecule whose amino acid sequence comprises a subset of the amino acid sequence of that protein. A protein or fragment thereof that comprises one or more additional peptide regions not derived from that protein is a "fusion" protein. Such molecules may be derivatized to contain carbohydrate or other moieties (such as keyhole limpet hemocyanin, *etc.*). Fusion protein or peptide molecule of the present invention are preferably produced via recombinant means.

Another class of agents comprise protein or peptide molecules encoded by SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof or, fragments or fusions thereof in which conservative, non-essential, or not relevant, amino acid residues have

been added, replaced, or deleted. An example of such a homologue is the homologue protein of all non-*Arabidopsis* plant species, including but not limited to alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, maize, an ornamental plant, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus etc.* Particularly preferred non-*Arabidopsis* plants to utilize for the isolation of homologues would include *Brassicaceae*. Such a homologue can be obtained by any of a variety of methods. For example, as indicated above, one or more of the disclosed sequences (SEQ ID NO: 1 through SEQ ID NO: 27,741 or complements thereof or fragments of either) will be used to define a pair of primers that may be used to isolate the homologue-encoding nucleic acid molecules from any desired species. Such molecules can be expressed to yield homologues by recombinant means.

(c) Antibodies

One aspect of the present invention concerns antibodies, single-chain antigen binding molecules, or other proteins that specifically bind to one or more of the protein or peptide molecules of the present invention and their homologues, fusions or fragments. Such antibodies may be used to quantitatively or qualitatively detect the protein or peptide molecules of the present invention. As used herein, an antibody or peptide is said to "specifically bind" to a protein or peptide molecule of the present invention if such binding is not competitively inhibited by the presence of non-related molecules. In a preferred embodiment the antibodies of the present invention bind to proteins of the present invention, in a more preferred embodiment of the antibodies of the present invention bind to a protein derived from *Arabidopsis* and more preferably bind to a protein or fragments thereof of *Arabidopsis thaliana, Columbia.*

Nucleic acid molecules that encode all or part of the protein of the present invention can be expressed, via recombinant means, to yield protein or peptides that can

in turn be used to elicit antibodies that are capable of binding the expressed protein or peptide. Such antibodies may be used in immunoassays for that protein. Such protein-encoding molecules, or their fragments may be a "fusion" molecule (*i.e.*, a part of a larger nucleic acid molecule) such that, upon expression, a fusion protein is produced. It is understood that any of the nucleic acid molecules of the present invention may be expressed, via recombinant means, to yield proteins or peptides encoded by these nucleic acid molecules.

The antibodies that specifically bind proteins and protein fragments of the present invention may be polyclonal or monoclonal, and may comprise intact immunoglobulins, or antigen binding portions of immunoglobulins (such as (F(ab'), $F(ab')_2$ fragments), or single-chain immunoglobulins producible, for example, via recombinant means. It is understood that practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of antibodies (see, for example, Harlow and Lane, *Antibodies: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1988), the entirety of which is herein incorporated by reference).

Murine monoclonal antibodies are particularly preferred. BALB/c mice are preferred for this purpose, however, equivalent strains may also be used. The animals are preferably immunized with approximately 25 µg of purified protein (or fragment thereof) that has been emulsified a suitable adjuvant (such as TiterMax adjuvant (Vaxcel, Norcross, GA)). Immunization is preferably conducted at two intramuscular sites, one intraperitoneal site, and one subcutaneous site at the base of the tail. An additional i.v. injection of approximately 25 µg of antigen is preferably given in normal saline three weeks later. After approximately 11 days following the second injection, the mice may be bled and the blood screened for the presence of anti-protein or peptide antibodies. Preferably, a direct binding Enzyme-Linked Immunoassay (ELISA) is employed for this purpose.

More preferably, the mouse having the highest antibody titer is given a third i.v. injection of approximately 25 μg of the same protein or fragment. The splenic leukocytes from this animal may be recovered 3 days later, and are then permitted to fuse, most preferably, using polyethylene glycol, with cells of a suitable myeloma cell line (such as, for example, the P3X63Ag8.653 myeloma cell line). Hybridoma cells are selected by culturing the cells under "HAT" (hypoxanthine-aminopterin-thymine) selection for about one week. The resulting clones may then be screened for their capacity to produce monoclonal antibodies ("mAbs"), preferably by direct ELISA.

In one embodiment, anti-protein or peptide monoclonal antibodies are isolated using a fusion of a protein, protein fragment, or peptide of the present invention, or conjugate of a protein, protein fragment, or peptide of the present invention, as immunogens. Thus, for example, a group of mice can be immunized using a fusion protein emulsified in Freund's complete adjuvant (*e.g.*, approximately 50 μg of antigen per immunization). At three week intervals, an identical amount of antigen is emulsified in Freund's incomplete adjuvant and used to immunize the animals. Ten days following the third immunization, serum samples are taken and evaluated for the presence of antibody. If antibody titers are too low, a fourth booster can be employed. Polysera capable of binding the protein or peptide can also be obtained using this method.

In a preferred procedure for obtaining monoclonal antibodies, the spleens of the above-described immunized mice are removed, disrupted, and immune splenocytes are isolated over a ficoll gradient. The isolated splenocytes are fused, using polyethylene glycol with BALB/c-derived HGPRT (hypoxanthine guanine phosphoribosyl transferase) deficient P3x63xAg8.653 plasmacytoma cells. The fused cells are plated into 96-well microtiter plates and screened for hybridoma fusion cells by their capacity to grow in culture medium supplemented with hypothanthine, aminopterin and thymidine for approximately 2-3 weeks.

Hybridoma cells that arise from such incubation are preferably screened for their capacity to produce an immunoglobulin that binds to a protein of interest. An indirect ELISA may be used for this purpose. In brief, the supernatants of hybridomas are incubated in microtiter wells that contain immobilized protein. After washing, the titer of bound immunoglobulin can be determined using, for example, a goat anti-mouse antibody conjugated to horseradish peroxidase. After additional washing, the amount of immobilized enzyme is determined (for example through the use of a chromogenic substrate). Such screening is performed as quickly as possible after the identification of the hybridoma in order to ensure that a desired clone is not overgrown by non-secreting neighbors. Desirably, the fusion plates are screened several times since the rates of hybridoma growth vary. In a preferred sub-embodiment, a different antigenic form of immunogen may be used to screen the hybridoma. Thus, for example, the splenocytes may be immunized with one immunogen, but the resulting hybridomas can be screened using a different immunogen. It is understood that any of the protein or peptide molecules of the present invention may be used to raise antibodies.

As discussed below, such antibody molecules or their fragments may be used for diagnostic purposes. Where the antibodies are intended for diagnostic purposes, it may be desirable to derivatize them, for example with a ligand group (such as biotin) or a detectable marker group (such as a fluorescent group, a radioisotope or an enzyme).

The ability to produce antibodies that bind the protein or peptide molecules of the present invention permits the identification of mimetic compounds of those molecules. A "mimetic compound" is a compound that is not that compound, or a fragment of that compound, but which nonetheless exhibits an ability to specifically bind to antibodies directed against that compound.

It is understood that any of the agents of the present invention can be substantially purified and/or be biologically active and/or recombinant

Uses of the Agents of the Invention

Nucleic acid molecules and fragments thereof of the present invention may be employed to obtain other nucleic acid molecules from the same species. Such nucleic acid molecules include the nucleic acid molecules that have the complete coding sequence of a protein and promoters and flanking sequences of such molecules. In addition, such nucleic acid molecules include nucleic acid molecules that encode for other isozymes or gene family members. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries obtained from *Arabidopsis*. Methods for forming such libraries are well known in the art.

Nucleic acid molecules and fragments thereof of the present invention may also be employed to obtain nucleic acid homologues. Such homologues include the nucleic acid molecules of other plants or other organisms (*e.g.*, alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, maize, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus, etc.*) including the nucleic acid molecules that encode, in whole or in part, protein homologues of other plant species or other organisms, sequences of genetic elements such as promoters and transcriptional regulatory elements. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries obtained from such plant species. Methods for forming such libraries are well known in the art. Such homologue molecules may differ in their nucleotide sequences from those found in one or more of SEQ ID NO:1 through SEQ ID NO: 51,470 or complements thereof because complete complementarity is not needed for stable hybridization. The nucleic acid molecules of the present invention therefore also include molecules that, although capable of specifically hybridizing with the nucleic acid

molecules may lack "complete complementarity." In a particular embodiment, methods or 3' or 5' RACE may be used (Frohman, M.A. *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:8998-9002 (1988); Ohara, O. *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:5673-5677 (1989)) to obtain such sequences.

Any of a variety of methods may be used to obtain one or more of the above-described nucleic acid molecules (Zamechik *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 83*:4143-4146 (1986), the entirety of which is herein incorporated by reference; Goodchild *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:5507-5511 (1988), the entirety of which is herein incorporated by reference; Wickstrom *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 85*:1028-1032 (1988), the entirety of which is herein incorporated by reference; Holt *et al.*, *Molec. Cell. Biol. 8*:963-973 (1988), the entirety of which is herein incorporated by reference; Gerwirtz *et al.*, *Science 242*:1303-1306 (1988), the entirety of which is herein incorporated by reference; Anfossi *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:3379-3383 (1989), the entirety of which is herein incorporated by reference; Becker *et al.*, *EMBO J. 8*:3685-3691 (1989); the entirety of which is herein incorporated by reference). Automated nucleic acid synthesizers may be employed for this purpose. In lieu of such synthesis, the disclosed nucleic acid molecules may be used to define a pair of primers that can be used with the polymerase chain reaction (Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol. 51*:263-273 (1986); Erlich *et al.*, European Patent 50,424; European Patent 84,796, European Patent 258,017, European Patent 237,362; Mullis, European Patent 201,184; Mullis *et al.*, U.S. Patent 4,683,202; Erlich, U.S. Patent 4,582,788; and Saiki, R. *et al.*, U.S. Patent 4,683,194, all of which are herein incorporated by reference in their entirety) to amplify and obtain any desired nucleic acid molecule or fragment.

Nucleic acid molecules and fragments thereof of the present invention may be employed for genetic mapping studies using linkage analysis (genetic markers). A genetic linkage map shows the relative locations of specific DNA markers along a

chromosome. Maps are used for the identification of genes associated with genetic diseases or phenotypic traits, comparative genomics, and as a guide for physical mapping. Through genetic mapping, a fine scale linkage map can be developed using DNA markers, and, then, a genomic DNA library of large-sized fragments can be screened with molecular markers linked to the desired trait. In a preferred embodiment of the present invention, the genomic library screen with the nucleic acid molecules of the present invention is a genomic library of *Arabidopsis*.

Mapping marker locations is based on the observation that two markers located near each other on the same chromosome will tend to be passed together from parent to offspring. During gamete production, DNA strands occasionally break and rejoin in different places on the same chromosome or on the homologous chromosome. The closer the markers are to each other, the more tightly linked and the less likely a recombination event will fall between and separate them. Recombination frequency thus provides an estimate of the distance between two markers.

In segregating populations, target genes have been reported to have been placed within an interval of 5-10 cM with a high degree of certainty (Tanksley *et al., Trends in Genetics 11(2):*63-68 (1995), the entirety of which is herein incorporated by reference). The markers defining this interval are used to screen a larger segregating population to identify individuals derived from one or more gametes containing a crossover in the given interval. Such individuals are useful in orienting other markers closer to the target gene. Once identified, these individuals can be analyzed in relation to all molecular markers within the region to identify those closest to the target.

Markers of the present invention can be employed to construct linkage maps and to locate genes with qualitative and quantitative effects. The genetic linkage of additional marker molecules can be established by a genetic mapping model such as, without limitation, the flanking marker model reported by Lander and Botstein, *Genetics 121*:185-199 (1989), and the interval mapping, based on maximum likelihood methods described

by Lander and Botstein, *Genetics 121*:185-199 (1989), the entirety of which is herein incorporated by reference and implemented in the software package MAPMAKER/QTL (Lincoln and Lander, *Mapping Genes Controlling Quantitative Traits Using MAPMAKER/QTL,* Whitehead Institute for Biomedical Research, Massachusetts, (1990)). Additional software includes Qgene, Version 2.23 (1996), Department of Plant Breeding and Biometry, 266 Emerson Hall, Cornell University, Ithaca, NY, the manual of which is herein incorporated by reference in its entirety). Use of the Qgene software is a particularly preferred approach.

A maximum likelihood estimate (MLE) for the presence of a marker is calculated, together with an MLE assuming no QTL effect, to avoid false positives. A $\log_{10}$ of an odds ratio (LOD) is then calculated as: LOD = $\log_{10}$ (MLE for the presence of a QTL/MLE given no linked QTL).

The LOD score essentially indicates how much more likely the data are to have arisen assuming the presence of a QTL than in its absence. The LOD threshold value for avoiding a false positive with a given confidence, say 95%, depends on the number of markers and the length of the genome. Graphs indicating LOD thresholds are set forth in Lander and Botstein, *Genetics 121*:185-199 (1989), the entirety of which is herein incorporated by reference and further described by Arús and Moreno-González, *Plant Breeding*, Hayward, Bosemark, Romagosa (eds.) Chapman & Hall, London, pp. 314-331 (1993).

Additional models can be used. Many modifications and alternative approaches to interval mapping have been reported, including the use of non-parametric methods (Kruglyak and Lander, *Genetics, 139:*1421-1428 (1995), the entirety of which is herein incorporated by reference). Multiple regression methods or models can be also used, in which the trait is regressed on a large number of markers (Jansen, *Biometrics in Plant Breed*, van Oijen, Jansen (eds.) Proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding, The Netherlands, pp. 116-124 (1994); Weber and Wricke,

Advances in Plant Breeding, Blackwell, Berlin, 16 (1994). Procedures combining interval mapping with regression analysis, where the phenotype is regressed onto a single putative QTL at a given marker interval, and at the same time onto a number of markers that serve as 'cofactors,' have been reported by Jansen and Stam, *Genetics, 136*:1447-1455 (1994) and Zeng, *Genetics, 136*:1457-1468 (1994). Generally, the use of cofactors reduces the bias and sampling error of the estimated QTL positions (Utz and Melchinger, *Biometrics in Plant Breeding,* van Oijen, Jansen (eds.) Proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding, The Netherlands, pp.195-204 (1994), thereby improving the precision and efficiency of QTL mapping (Zeng, *Genetics, 136*:1457-1468 (1994). These models can be extended to multi-environment experiments to analyze genotype-environment interactions (Jansen *et al., Theo. Appl. Genet. 91*:33-37 (1995).

Selection of an appropriate mapping populations is important to map construction. The choice of appropriate mapping population depends on the type of marker systems employed (Tanksley *et al., J.P. Gustafson and R. Appels* (eds.). Plenum Press, New York, pp. 157-173 (1988), the entirety of which is herein incorporated by reference). Consideration must be given to the source of parents (adapted vs. exotic) used in the mapping population. Chromosome pairing and recombination rates can be severely disturbed (suppressed) in wide crosses (adapted x exotic) and generally yield greatly reduced linkage distances. Wide crosses will usually provide segregating populations with a relatively large array of polymorphisms when compared to progeny in a narrow cross (adapted x adapted).

An F_2 population is the first generation of selfing after the hybrid seed is produced. Usually a single F_1 plant is selfed to generate a population segregating for all the genes in Mendelian (1:2:1) fashion. Maximum genetic information is obtained from a completely classified F_2 population using a codominant marker system (Mather, Measurement of Linkage in Heredity: Methuen and Co., (1938), the entirety of which is

herein incorporated by reference). In the case of dominant markers, progeny tests (*e.g.*, F_3, BCF_2) are required to identify the heterozygotes, thus making it equivalent to a completely classified F_2 population. However, this procedure is often prohibitive because of the cost and time involved in progeny testing. Progeny testing of F_2 individuals is often used in map construction where phenotypes do not consistently reflect genotype (*e.g.*, disease resistance) or where trait expression is controlled by a QTL. Segregation data from progeny test populations *e.g.*, F_3 or BCF_2) can be used in map construction. Marker-assisted selection can then be applied to cross progeny based on marker-trait map associations (F_2, F_3), where linkage groups have not been completely disassociated by recombination events (*i.e.*, maximum disequilibrium).

Recombinant inbred lines (RIL) (genetically related lines; usually >F_5, developed from continuously selfing F_2 lines towards homozygosity) can be used as a mapping population. Information obtained from dominant markers can be maximized by using RIL because all loci are homozygous or nearly so. Under conditions of tight linkage (*i.e.*, about <10% recombination), dominant and co-dominant markers evaluated in RIL populations provide more information per individual than either marker type in backcross populations (Reiter. *Proc. Natl. Acad. Sci. (U.S.A.) 89*:1477-1481 (1992), the entirety of which is herein incorporated by reference). However, as the distance between markers becomes larger (*i.e.*, loci become more independent), the information in RIL populations decreases dramatically when compared to codominant markers.

Backcross populations (*e.g.*, generated from a cross between a successful variety (recurrent parent) and another variety (donor parent) carrying a trait not present in the former) can be utilized as a mapping population. A series of backcrosses to the recurrent parent can be made to recover most of its desirable traits. Thus a population is created consisting of individuals nearly like the recurrent parent but each individual carries varying amounts or mosaic of genomic regions from the donor parent. Backcross populations can be useful for mapping dominant markers if all loci in the recurrent parent

are homozygous and the donor and recurrent parent have contrasting polymorphic marker alleles (Reiter *et al., Proc. Natl. Acad. Sci. (U.S.A. 89*:1477-1481 (1992)). Information obtained from backcross populations using either codominant or dominant makers is less than that obtained from F_2 populations because one, rather than two, recombinant gametes are sampled per plant. Backcross populations, however, are more informative (at low marker saturation) when compared to RILs as the distance between linked loci increases in RIL populations (*i.e.*, about .15% recombination). Increased recombination can be beneficial for resolution of tight linkages, but may be undesirable in the construction of maps with low marker saturation.

Near-isogenic lines (NIL)(created by many backcrosses to produce an array of individuals that are nearly identical in genetic composition except for the trait or genomic region under interrogation) can be used as a mapping population. In mapping with NILs, only a portion of the polymorphic loci are expected to map to a selected region.

Bulk segregant analysis (BSA) is a method developed for the rapid identification of linkage between markers and traits of interest (Michelmore *et al., Proc. Natl. Acad. Sci. U.S.A. 88*:9828-9832 (1991). In BSA, two bulked DNA samples are drawn from a segregating population originating from a single cross. These bulks contain individuals that are identical for a particular trait (resistant or susceptible to particular disease) or genomic region but arbitrary at unlinked regions (*i.e.*, heterozygous). Regions unlinked to the target region will not differ between the bulked samples of many individuals in BSA.

Applications for markers in plant breeding include: Quantitative Trait Loci (QTL) mapping (Edwards *et al., Genetics* 116:113-115 (1987), the entirety of which is herein incorporated by reference); Nienhuis *et al., Crop Sci. 27*:797-803 (1987); Osborn *et al., Theor. Appl. Genet. 73*:350-356 (1987); Romero-Severson *et al., Use of RFLPs In Analysis of Quantitative Trait Loci In* Maize, In Helentjaris and Burr (eds.) pp. 97-102 (1989), the entirety of which is herein incorporated by reference; Young *et al., Genetics*

120:570-585 (1988), the entirety of which is herein incorporated by reference; Martin *et al., Science 243*:1725-1728 (1989), the entirety of which is herein incorporated by reference): Sarfatti *et al. Theor. Appl Genet. 78*:22-26 (1989), the entirety of which is herein incorporated by reference; Tanksley *et al., Biotech. 7*:257-264 (1989); Barone *et al., Mol. Gen. Genet. 224*:177-182 (1990), the entirety of which is herein incorporated by reference); Jung *et al., Theor, Appl. Genet. 79*:663-672 (1990), the entirety of which is herein incorporated by reference; Keim *et al., Genetics 126*:735-742 (1990), the entirety of which is herein incorporated by reference, *Theor. Appl. Genet. 79*:465-369 (1990), the entirety of which is herein incorporated by reference; Paterson *et al., Genetics 124*:735-742 (1990), the entirety of which is herein incorporated by reference; Martin *et al., Proc. Natl. Acad. Sci. (U.S.A.) 88*:2336-2340 (1991), the entirety of which is herein incorporated by reference; Messeguer *et al., Theor. Appl. Genet. 82*:529-536 (1991), the entirety of which is herein incorporated by reference; Michelmore *et al., Proc Natl. Acad. Sci. (U.S.A.) 88*:9828-9832 (1991), the entirety of which is herein incorporated by reference; Ottaviano *et al., Theor. Appl. Genet. 81*:713-719 (1991), the entirety of which is herein incorporated by reference; Yu *et al., Theor. Appl. Genet. 81*:471-476 (1991), the entirety of which is herein incorporated by reference; Diers *et al., Crop Sci. 32*:77-383 (1992), the entirety of which is herein incorporated by reference, *Theor. Appl. Genet. 83*:608-612 (1992), the entirety of which is herein incorporated by reference, *J. Plant Nut.* 15:2127-2136 (1992), the entirety of which is herein incorporated by reference; Doebley *et al., Proc. Natl. Acad. Sci. (U.S.A.) 87*:9888-9892 (1990), the entirety of which is herein incorporated by reference), screening genetic resource strains for useful quantitative trait alleles and introgression of these alleles into commercial varieties (Beckmann and Soller, *Theor. Appl. Genet.* 67:35-43 (1983), the entirety of which is herein incorporated by reference; Tanksley *et al.*, (1989) the entirety of which is incorporated by reference), or the mapping of mutations (Rafalski *et al.*, In: *Nonmammalian Genomic Analysis*, ed. Birren and Lai, Academic Press, San Diego, CA, pp. 75-134 (1996). Additionally,

markers can be used to characterize transformants or germplasm, as a genetic diagnostic test for plant breeding or to identify individuals or varieties (Soller and Beckmann, *Theor. Appl. Genet.* (67):25-33 (1983), the entirety of which is herein incorporated by reference; Tanksley *et al.*, (1989). Markers also can be used to obtain information about: (1) the number, effect, and chromosomal location of each gene affecting a trait; (2) effects of multiple copies of individual genes (gene dosage); (3) interaction between/among genes controlling a trait (epistasis); (4) whether individual genes affect more than one trait (pleiotropy); and (5) stability of gene function across environments (Gx E interactions).

It is understood that one or more of the nucleic acid molecules of the present invention may in one embodiment be used as markers in genetic mapping. In a preferred embodiment, nucleic acid molecules of the present invention may in one embodiment be used as markers with non-*Arabidopsis* plant species, including but not limited to alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, maize, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus etc.* Particularly preferred non-*Arabidopsis* plants are *Brassicacae.*

The nucleic acid molecules of the present invention may be used for physical mapping. Physical mapping, in conjunction with linkage analysis, can enable the isolation of genes. Physical mapping has been reported to identify the markers closest in terms of genetic recombination to a gene target for cloning. Once a DNA marker is linked to a gene of interest, the chromosome walking technique can be used to find the genes via overlapping clones. For chromosome walking, random molecular markers or established molecular linkage maps are used to conduct a search to localize the gene adjacent to one or more markers. A chromosome walk is then initiated from the closest linked marker (Bukanov and Berg, *Mo. Microbiol. 11:*509-523 (1994), the entirety of which is herein incorporated by reference; Birkenbihl and Vielmetter, *Nucleic Acids Res.*

*17:*5057-5069 (1989), the entirety of which is herein incorporated by reference; Wenzel and Herrmann, *Nucleic Acids Res. 16:*8323-8336, (1988), the entirety of which is herein incorporated by reference). Starting from the selected clones, labeled probes specific for the ends of the insert DNA are synthesized and used as probes in hybridizations against a representative library. Clones hybridizing with one of the probes are picked and serve as templates for the synthesis of new probes; by subsequent analysis, contigs are produced.

The degree of overlap of the hybridizing clones used to produce a contig can be determined by comparative restriction analysis. Comparative restriction analysis can be carried out in different ways all of which exploit the same principle; two clones of a library are very likely to overlap if they contain a limited number of restriction sites for one or more restriction endonucleases located at the same distance from each other. The most frequently used procedures are, fingerprinting (Coulson *et al., Proc. Natl. Acad. Sci. (U.S.A.) 83:*7821-7821, (1986), the entirety of which is herein incorporated by reference; Knott *et al., Nucleic Acids Res. 16:*2601-2612 (1988), the entirety of which is herein incorporated by reference; Eiglmeier *et al. Mol. Microbiol.* 7:197-206 (1993), the entirety of which is herein incorporated by reference, (1993), restriction fragment mapping (Smith and Birnstiel, *Nucleic Acids Res.* 3:2387-2398 (1976), the entirety of which is herein incorporated by reference, or the "landmarking" technique (Charlebois *et al. J. Mol. Biol.* 222:509-524 (1991), the entirety of which is herein incorporated by reference

It is understood that the nucleic acid molecules of the present invention may in one embodiment be used in physical mapping. In a preferred embodiment, nucleic acid molecules of the present invention may in one embodiment be used in the physical mapping of *Brassicaceae,* particularly *Arabidopsis.*

Nucleic acid molecules of the present invention can be used in comparative mapping. Comparative mapping within families provides a method to assess the degree of sequence conservation, gene order, ploidy of species, ancestral relationships and the rates at which individual genomes are evolving. Comparative mapping has been carried

out by cross-hybridizing molecular markers across species within a given family. As in genetic mapping, molecular markers are needed but instead of direct hybridization to mapping filters, the markers are used to select large insert clones from a total genomic DNA library of a related species. The selected clones, each a representative of a single marker, can then be used to physically map the region in the target species. The advantage of this method for comparative mapping is that no mapping population or linkage map of the target species is needed and the clones may also be used in other closely related species. By comparing the results obtained by genetic mapping in model plants, with those from other species, similarities of genomic structure among plants species can be established. Cross-hybridization of RFLP markers have been reported and conserved gene order has been established in many studies. Such macroscopic synteny is utilized for the estimation of correspondence of loci among these crops. These loci include not only Mendelian genes but also Quantitative Trait Loci (QTLs) (Mohan *et al.*, *Molecular Breeding* 3:87-103 (1997), the entirety of which is herein incorporated by reference. It is understood that markers of the present invention may in another embodiment be used in comparative mapping. In a preferred embodiment the markers of present invention may be used in the comparative mapping of non-*Arabidopsis* plant species, including but not limited to alfalfa, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, an ornamental plant, maize, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus etc.* Particularly preferred non-*Arabidopsis* plants to utilize for comparative mapping are the *Brassicaceae*.

The nucleic acid molecules of the present invention can be used to identify polymorphisms. In one embodiment of the present invention, a nucleic acid molecule of the present invention (or a sub-fragment of either) may be employed as a marker nucleic acid molecule to identify such polymorphism(s). Alternatively, such polymorphisms can

be detected through the use of a marker nucleic acid molecule or a marker protein that is genetically linked to (*i.e.*, a polynucleotide that co-segregates with) such polymorphism(s).

In an alternative embodiment, such polymorphisms can be detected through the use of a marker nucleic acid molecule that is physically linked to such polymorphism(s). For this purpose, marker nucleic acid molecules comprising a nucleotide sequence of a polynucleotide located within 1 mb of the polymorphism(s), and more preferably within 100 kb of the polymorphism(s), and most preferably within 10 kb of the polymorphism(s) can be employed.

The genomes of animals and plants naturally undergo spontaneous mutation in the course of their continuing evolution (Gusella, *Ann. Rev. Biochem. 55*:831-854 (1986)). A "polymorphism" is a variation or difference in the sequence of the gene or its flanking regions that arises in some of the members of a species. The variant sequence and the "original" sequence co-exist in the species' population. In some instances, such co-existence is in stable or quasi-stable equilibrium.

A polymorphism is thus said to be "allelic," in that, due to the existence of the polymorphism, some members of a species may have the original sequence (*i.e.*, the original "allele") whereas other members may have the variant sequence (*i.e.*, the variant "allele"). In the simplest case, only one variant sequence may exist, and the polymorphism is thus said to be di-allelic. In other cases, the species' population may contain multiple alleles, and the polymorphism is termed tri-allelic, *etc.* A single gene may have multiple different unrelated polymorphisms. For example, it may have a di-allelic polymorphism at one site, and a multi-allelic polymorphism at another site.

The variation that defines the polymorphism may range from a single nucleotide variation to the insertion or deletion of extended regions within a gene. In some cases, the DNA sequence variations are in regions of the genome that are characterized by short tandem repeats (STRs) that include tandem di- or tri-nucleotide repeated motifs of

nucleotides. Polymorphisms characterized by such tandem repeats are referred to as "variable number tandem repeat" ("VNTR") polymorphisms. VNTRs have been used in identity analysis (Weber, U.S. Patent 5,075,217; Armour *et al.*, *FEBS Lett. 307*:113-115 (1992); Jones *et al.*, *Eur. J. Haematol. 39*:144-147 (1987); Horn *et al.*, PCT Application WO91/14003; Jeffreys, European Patent Application 370,719; Jeffreys, U.S. Patent 5,175,082; Jeffreys *et al.*, *Amer. J. Hum. Genet. 39*:11-24 (1986); Jeffreys *et al.*, *Nature 316*:76-79 (1985); Gray *et al.*, *Proc. R. Acad. Soc. Lond. 243*:241-253 (1991); Moore *et al.*, *Genomics 10*:654-660 (1991); Jeffreys *et al.*, *Anim. Genet. 18*:1-15 (1987); Hillel *et al.*, *Anim. Genet. 20*:145-155 (1989); Hillel *et al.*, *Genet. 124*:783-789 (1990), all of which are herein incorporated by reference in their entirety).

The detection of polymorphic sites in a sample of DNA may be facilitated through the use of nucleic acid amplification methods. Such methods specifically increase the concentration of polynucleotides that span the polymorphic site, or include that site and sequences located either distal or proximal to it. Such amplified molecules can be readily detected by gel electrophoresis or other means.

The most preferred method of achieving such amplification employs the polymerase chain reaction ("PCR") (Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol. 51*:263-273 (1986); Erlich *et al.*, European Patent Appln. 50,424; European Patent Appln. 84,796, European Patent Application 258,017, European Patent Appln. 237,362; Mullis, European Patent Appln. 201,184; Mullis *et al.*, U.S. Patent No. 4,683,202; Erlich., U.S. Patent No. 4,582,788; and Saiki *et al.*, U.S. Patent No. 4,683,194, all of which are herein incorporated by reference), using primer pairs that are capable of hybridizing to the proximal sequences that define a polymorphism in its double-stranded form.

In lieu of PCR, alternative methods, such as the "Ligase Chain Reaction" ("LCR") may be used (Barany, *Proc. Natl. Acad. Sci. (U.S.A.) 88*:189-193 (1991), the entirety of which is herein incorporated by reference. LCR uses two pairs of oligonucleotide probes

to exponentially amplify a specific target. The sequences of each pair of oligonucleotides is selected to permit the pair to hybridize to abutting sequences of the same strand of the target. Such hybridization forms a substrate for a template-dependent ligase. As with PCR, the resulting products thus serve as a template in subsequent cycles and an exponential amplification of the desired sequence is obtained.

LCR can be performed with oligonucleotides having the proximal and distal sequences of the same strand of a polymorphic site. In one embodiment, either oligonucleotide will be designed to include the actual polymorphic site of the polymorphism. In such an embodiment, the reaction conditions are selected such that the oligonucleotides can be ligated together only if the target molecule either contains or lacks the specific nucleotide that is complementary to the polymorphic site present on the oligonucleotide. Alternatively, the oligonucleotides may be selected such that they do not include the polymorphic site (see, Segev, PCT Application WO 90/01069, the entirety of which is herein incorporated by reference).

The "Oligonucleotide Ligation Assay" ("OLA") may alternatively be employed (Landegren *et al.*, *Science 241*:1077-1080 (1988), the entirety of which is herein incorporated by reference). The OLA protocol uses two oligonucleotides which are designed to be capable of hybridizing to abutting sequences of a single strand of a target. OLA, like LCR, is particularly suited for the detection of point mutations. Unlike LCR, however, OLA results in "linear" rather than exponential amplification of the target sequence.

Nickerson *et al.* have described a nucleic acid detection assay that combines attributes of PCR and OLA (Nickerson *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 87*:8923-8927 (1990), the entirety of which is herein incorporated by reference). In this method, PCR is used to achieve the exponential amplification of target DNA, which is then detected using OLA. In addition to requiring multiple, and separate, processing steps,

one problem associated with such combinations is that they inherit all of the problems associated with PCR and OLA.

Schemes based on ligation of two (or more) oligonucleotides in the presence of nucleic acid having the sequence of the resulting "di-oligonucleotide," thereby amplifying the di-oligonucleotide, are also known (Wu *et al.*, *Genomics 4*:560 (1989), the entirety of which is herein incorporated by reference), and may be readily adapted to the purposes of the present invention.

Other known nucleic acid amplification procedures, such as allele-specific oligomers, branched DNA technology, transcription-based amplification systems, or isothermal amplification methods may also be used to amplify and analyze such polymorphisms (Malek *et al.*, U.S. Patent 5,130,238; Davey *et al.*, European Patent Application 329,822; Schuster *et al.*, U.S. Patent 5,169,766; Miller *et al.*, PCT Application WO 89/06700; Kwoh *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:1173-1177 (1989); Gingeras *et al.*, PCT Application WO 88/10315; Walker *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 89*:392-396 (1992), all of which are herein incorporated by reference in their entirety).

The identification of a polymorphism can be determined in a variety of ways. By correlating the presence or absence of it in an plant with the presence or absence of a phenotype, it is possible to predict the phenotype of that plant. If a polymorphism creates or destroys a restriction endonuclease cleavage site, or if it results in the loss or insertion of DNA (*e.g.*, a VNTR polymorphism), it will alter the size or profile of the DNA fragments that are generated by digestion with that restriction endonuclease. As such, individuals that possess a variant sequence can be distinguished from those having the original sequence by restriction fragment analysis. Polymorphisms that can be identified in this manner are termed "restriction fragment length polymorphisms" ("RFLPs"). RFLPs have been widely used in human and plant genetic analyses (Glassberg, UK Patent Application 2135774; Skolnick *et al.*, *Cytogen. Cell Genet. 32*:58-67 (1982); Botstein *et*

al., *Ann. J. Hum. Genet. 32*:314-331 (1980); Fischer *et al.*, (PCT Application WO90/13668); Uhlen, PCT Application WO90/11369).

Polymorphisms can also be identified by Single Strand Conformation Polymorphism (SSCP) analysis. The SSCP technique is a method capable of identifying most sequence variations in a single strand of DNA, typically between 150 and 250 nucleotides in length (Elles, *Methods in Molecular Medicine: Molecular Diagnosis of Genetic Diseases*, Humana Press (1996), the entirety of which is herein incorporated by reference); Orita *et al., Genomics 5*:874-879 (1989), the entirety of which is herein incorporated by reference). Under denaturing conditions a single strand of DNA will adopt a conformation that is uniquely dependent on its sequence conformation. This conformation usually will be different, even if only a single base is changed. Most conformations have been reported to alter the physical configuration or size sufficiently to be detectable by electrophoresis. A number of protocols have been described for SSCP including, but not limited to Lee *et al., Anal. Biochem. 205*:289-293 (1992), the entirety of which is herein incorporated by reference; Suzuki *et al., Anal. Biochem. 192*:82-84 (1991), the entirety of which is herein incorporated by reference; Lo *et al., Nucleic Acids Research 20*:1005-1009 (1992), the entirety of which is herein incorporated by reference; Sarkar *et al., Genomics 13*:441-443 (1992), the entirety of which is herein incorporated by reference). It is understood that one or more of the nucleic acids of the present invention, may be utilized as markers or probes to detect polymorphisms by SSCP analysis.

Polymorphisms may also be found using a DNA fingerprinting technique called amplified fragment length polymorphism (AFLP), which is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA to profile that DNA. Vos *et al.*, *Nucleic Acids Res. 23*:4407-4414 (1995), the entirety of which is herein incorporated by reference. This method allows for the specific co-amplification of high

numbers of restriction fragments, which can be visualized by PCR without knowledge of the nucleic acid sequence.

AFLP employs basically three steps. Initially, a sample of genomic DNA is cut with restriction enzymes and oligonucleotide adapters are ligated to the restriction fragments of the DNA. The restriction fragments are then amplified using PCR by using the adapter and restriction sequence as target sites for primer annealing. The selective amplification is achieved by the use of primers that extend into the restriction fragments, amplifying only those fragments in which the primer extensions match the nucleotide flanking the restriction sites. These amplified fragments are then visualized on a denaturing polyacrylamide gel.

AFLP analysis has been performed on *Salix* (Beismann *et al.*, *Mol. Ecol. 6*:989-993 (1997), the entirety of which is herein incorporated by reference); *Acinetobacter* (Janssen *et al.*, *Int. J. Syst. Bacteriol 47*:1179-1187 (1997), the entirety of which is herein incorporated by reference), *Aeromonas popoffi* (Huys *et al.*, *Int. J. Syst. Bacteriol. 47*:1165-1171 (1997), the entirety of which is herein incorporated by reference), rice (McCouch *et al.*, *Plant Mol. Biol. 35*:89-99 (1997), the entirety of which is herein incorporated by reference); Nandi *et al.*, *Mol. Gen. Genet. 255*:1-8 (1997); Cho *et al.*, *Genome 39*:373-378 (1996), herein incorporated by reference), barley (*Hordeum vulgare*)(Simons *et al.*, *Genomics 44*:61-70 (1997), the entirety of which is herein incorporated by reference; Waugh *et al.*, *Mol. Gen. Genet. 255*:311-321 (1997), the entirety of which is herein incorporated by reference; Qi *et al.*, *Mol. Gen Genet. 254*:330-336 (1997), the entirety of which is herein incorporated by reference; Becker *et al.*, *Mol. Gen. Genet. 249*:65-73 (1995), the entirety of which is herein incorporated by reference), potato (Van der Voort *et al.*, *Mol. Gen. Genet. 255*:438-447 (1997), the entirety of which is herein incorporated by reference; Meksem *et al.*, *Mol. Gen. Genet. 249*:74-81 (1995), the entirety of which is herein incorporated by reference), *Phytophthora infestans* (Van der Lee *et al.*, *Fungal Genet. Biol. 21*:278-291 (1997), the entirety of which is herein

incorporated by reference), *Bacillus anthracis* (Keim *et al.*, *J. Bacteriol. 179*:818-824 (1997)), *Astragalus cremnophylax* (Travis *et al.*, *Mol. Ecol. 5*:735-745 (1996), the entirety of which is herein incorporated by reference), *Arabidopsis* (Cnops *et al.*, *Mol. Gen. Genet. 253*:32-41 (1996), the entirety of which is herein incorporated by reference), *Escherichia coli* (Lin *et al.*, *Nucleic Acids Res. 24*:3649-3650 (1996), the entirety of which is herein incorporated by reference), *Aeromonas* (Huys *et al.*, *Int. J. Syst. Bacteriol. 46*:572-580 (1996), the entirety of which is herein incorporated by reference), nematode (Folkertsma *et al.*, *Mol. Plant Microbe Interact. 9*:47-54 (1996), the entirety of which is herein incorporated by reference), tomato (Thomas *et al.*, *Plant J. 8*:785-794 (1995), the entirety of which is herein incorporated by reference), and human (Latorra *et al.*, *PCR Methods Appl. 3*:351-358 (1994)the entirety of which is herein incorporated by reference). AFLP analysis has also been used for fingerprinting mRNA (Money *et al.*, *Nucleic Acids Res. 24*:2616-2617 (1996), the entirety of which is herein incorporated by reference; Bachem *et al.*, *Plant J. 9*:745-753 (1996), the entirety of which is herein incorporated by reference). It is understood that one or more of the nucleic acid molecules of the present invention, may be utilized as markers or probes to detect polymorphisms by AFLP analysis for fingerprinting mRNA.

Polymorphisms may also be found using random amplified polymorphic DNA (RAPD) (Williams *et al.*, *Nucl. Acids Res. 18*:6531-6535 (1990), the entirety of which is herein incorporated by reference) and cleaveable amplified polymorphic sequences (CAPS) (Lyamichev *et al.*, *Science 260*:778-783 (1993), the entirety of which is herein incorporated by reference). It is understood that one or more of the nucleic acid molecules of the present invention, may be utilized as markers or probes to detect polymorphisms by RAPD or CAPS analysis.

Nucleic acid molecules of the present invention can be used to monitor expression. In an aspect of the present invention, one or more of the nucleic molecules of the present invention are used to determine whether a plant has a mutation affecting the

level (*i.e.*, the concentration of mRNA in a sample, *etc.*) or pattern (*i.e.*, the kinetics of expression, rate of decomposition, stability profile, *etc.*) of the expression encoded in part or whole by one or more of the nucleic acid molecule of the present invention (collectively, the "Expression Response" of a cell or tissue). As used herein, the Expression Response manifested by a cell or tissue is said to be "altered" if it differs from the Expression Response of cells or tissues of plants not exhibiting the phenotype. To determine whether a Expression Response is altered, the Expression Response manifested by the cell or tissue of the plant exhibiting the phenotype is compared with that of a similar cell or tissue sample of a plant not exhibiting the phenotype. As will be appreciated, it is not necessary to re-determine the Expression Response of the cell or tissue sample of plants not exhibiting the phenotype each time such a comparison is made; rather, the Expression Response of a particular plant may be compared with previously obtained values of normal plants. As used herein, the phenotype of the organism is any of one or more characteristics of an organism (*e.g.*, disease resistance, pest tolerance, environmental tolerance such as tolerance to abiotic stress, male sterility, quality improvement or yield *etc.*). A change in genotype or phenotype may be transient or permanent. Also as used herein, a tissue sample is any sample that comprises more than one cell. In a preferred aspect, a tissue sample comprises cells that share a common characteristic (*e.g.*, derived from root, seed, flower, leaf, stem or pollen *etc.*).

In one sub-aspect, such an analysis is conducted by determining the presence and/or identity of polymorphism(s) by one or more of the nucleic acid molecules of the present invention which are associated with a phenotype, or a predisposition to that phenotype.

A microarray-based method for high-throughput monitoring of plant gene expression may be utilized to measure gene-specific hybridization targets. This 'chip'-based approach involves using microarrays of nucleic acid molecules as gene-specific hybridization targets to quantitatively measure expression of the corresponding plant

genes (Schena *et al.*, *Science 270*:467-470 (1995), the entirety of which is herein incorporated by reference; Shalon, Ph.D. Thesis, Stanford University (1996), the entirety of which is herein incorporated by reference). Every nucleotide in a large sequence can be queried at the same time. Hybridization can be used to efficiently analyze nucleotide sequences.

Several microarray methods have been described. One method compares the sequences to be analyzed by hybridization to a set of oligonucleotides or cDNA molecules representing all possible subsequences (Bains and Smith, *J. Theor. Biol. 135*:303 (1989), the entirety of which is herein incorporated by reference). A second method hybridizes the sample to an array of oligonucleotide or cDNA probes. An array consisting of oligonucleotides or cDNA molecules complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Nucleic acid molecules microarrays may also be screened with protein molecules or fragments thereof to determine nucleic acid molecules that specifically bind protein molecules or fragments thereof.

The microarray approach may also be used with polypeptide targets (U.S. Patent No. 5,445,934; U.S. Patent No. 5,143,854; U.S. Patent No. 5,079,600; U.S. Patent No. 4,923,901, all of which are herein incorporated by reference in their entirety). Essentially, polypeptides are synthesized on a substrate (microarray) and these polypeptides can be screened with either protein molecules or fragments thereof or nucleic acid molecules in order to screen for either protein molecules or fragments thereof or nucleic acid molecules that specifically bind the target polypeptides (Fodor *et al.*, *Science 251*:767-773 (1991), the entirety of which is herein incorporated by reference).

It is understood that one or more of the molecules of the present invention, preferably one or more of the nucleic acid molecules or protein molecules or fragments thereof of the present invention may be utilized in a microarray based method. In a

preferred embodiment of the present invention, one or more of the *Arabidopsis Columbia* nucleic acid molecules or protein molecules or fragments thereof of the present invention may be utilized in a microarray based method. A particular preferred microarray embodiment of the present invention is a microarray comprising nucleic acid molecules encoding genes or fragments thereof that are homologues of known genes or nucleic acid molecules that comprise genes or fragments thereof that elicit only limited or no matches to known genes. A further preferred microarray embodiment of the present invention is a microarray comprising nucleic acid molecules encoding genes or fragments thereof that are homologues of known genes and nucleic acid molecules that comprise genes or fragments thereof that elicit only limited or no matches to known genes.

Nucleic acid molecules of the present invention may be used in site directed mutagenesis. Site-directed mutagenesis may be utilized to modify nucleic acid sequences, particularly as it is a technique that allows one or more of the amino acids encoded by a nucleic acid molecule to be altered (*e.g.*, a threonine to be replaced by a methionine). Three basic methods for site-directed mutagenesis are often employed. These are cassette mutagenesis (Wells *et al.*, *Gene 34*:315-23 (1985), the entirety of which is herein incorporated by reference), primer extension (Gilliam *et al.*, *Gene 12*:129-137 (1980), the entirety of which is herein incorporated by reference; Zoller and Smith, *Methods Enzymol. 100*:468-500 (1983), the entirety of which is herein incorporated by reference; Dalbadie-McFarland *et al., Proc. Natl. Acad. Sci. (U.S.A.) 79*:6409-6413 (1982), the entirety of which is herein incorporated by reference) and methods based upon PCR (Scharf *et al., Science 233*:1076-1078 (1986), the entirety of which is herein incorporated by reference; Higuchi *et al., Nucleic Acids Res. 16*:7351-7367 (1988), the entirety of which is herein incorporated by reference). Site-directed mutagenesis approaches are also described in European Patent 0 385 962, the entirety of which is herein incorporated by reference, European Patent 0 359 472, the entirety of which is

herein incorporated by reference, and U.S. Patent No. 5,625,136, the entirety of which is herein incorporated by reference.

Site-directed mutagenesis strategies have been applied to plants for both *in vitro* as well as *in vivo* site-directed mutagenesis (Lanz *et al., J. Biol. Chem. 266*:9971-9976 (1991), the entirety of which is herein incorporated by reference; Kovgan and Zhdanov, *Biotekhnologiya 5*:148-154, No. 207160n, Chemical Abstracts 110:225 (1989), the entirety of which is herein incorporated by reference; Ge *et al., Proc. Natl. Acad. Sci. (U.S.A.) 86*:4037-4041 (1989), the entirety of which is herein incorporated by reference; Zhu *et al., J. Biol. Chem. 271*:18494-18498 (1996), Chu *et al., Biochemistry 33*:6150-6157 (1994), the entirety of which is herein incorporated by reference; Small *et al., EMBO J. 11*:1291-1296 (1992), the entirety of which is herein incorporated by reference; Cho *et al., Mol. Biotechnol. 8*:13-16 (1997), Kita *et al., J. Biol. Chem. 271*:26529-26535 (1996), the entirety of which is herein incorporated by reference, Jin *et al., Mol. Microbiol. 7*:555-562 (1993), the entirety of which is herein incorporated by reference, Hatfield and Vierstra, *J. Biol. Chem. 267*:14799-14803 (1992), the entirety of which is herein incorporated by reference, Zhao *et al., Biochemistry 31*:5093-5099 (1992), the entirety of which is herein incorporated by reference).

Any of the nucleic acid molecules of the present invention may either be modified by site-directed mutagenesis or used as, for example, nucleic acid molecules that are used to target other nucleic acid molecules for modification. It is understood that mutants with more than one altered nucleotide can be constructed using techniques that practitioners skilled in the art are familiar with such as isolating restriction fragments and ligating such fragments into an expression vector (*see*, for example, Sambrook *et al., Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, New York (1989)). In a preferred embodiment of the present invention, one or more of the *Arabidopsis* nucleic acid molecules or fragments thereof of the present invention may be modified by site-directed mutagenesis.

Nucleic acid molecules of the present invention may be used in transformation. Exogenous genetic material may be transferred into a plant cell and the plant cell regenerated into a whole, fertile or sterile plant. Exogenous genetic material is any genetic material, whether naturally occurring or otherwise, from any source that is capable of being inserted into any organism. In a preferred embodiment of the present invention the exogenous genetic material can include *Arabidopsis* genetic material. Such genetic material may be transferred into either monocotyledons and dicotyledons including, but not limited to maize (pp. 63-69), soybean (pp. 50-60), *Arabidopsis* (p 45), phaseolus (pp. 47-49), peanut (pp. 49-50), alfalfa (p 60), wheat (pp. 69-71), rice (pp. 72-79), oat (pp. 80-81), sorghum (p 83), rye (p 84), tritordeum (p 84), millet (p85), fescue (p 85), perennial ryegrass (p 86), sugarcane (p87), cranberry (p101), papaya (pp. 101-102), banana (p 103), banana (p 103), muskmelon (p 104), apple (p 104), cucumber (p 105), dendrobium (p 109), gladiolus (p 110), chrysanthemum (p 110), liliacea (p 111), cotton (pp113-114), eucalyptus (p 115), sunflower (p 118), canola (p 118), turfgrass (p121), sugarbeet (p 122), coffee (p 122), and dioscorea (p 122), (Christou, In: *Particle Bombardment for Genetic Engineering of Plants,* Biotechnology Intelligence Unit. Academic Press, San Diego, California (1996), the entirety of which is herein incorporated by reference). Transfer of a nucleic acid that encodes for a protein can result in overexpression of that protein in a transformed cell or transgenic plant. One or more of the proteins or fragments thereof encoded by nucleic acid molecules of the present invention may be overexpressed in a transformed cell or transformed plant. Such overexpression may be the result of transient or stable transfer of the exogenous material.

Exogenous genetic material may be transferred into a plant cell by the use of a DNA vector or construct designed for such a purpose. Vectors have been engineered for transformation of large DNA inserts into plant genomes. BACs (binary bacterial artificial chromosomes) have been designed to replicate in both *E. coli* and *A. tumefaciens* and have all of the features required for transferring large inserts of DNA into plant

chromosomes Choi and Wing, http://genome.clemson.edu/protocols2-nj.html July, 1998. ApBACwich system has been developed to achieve site-directed integration of DNA into the genome. A 150 kb cotton BAC DNA is reported to have been transferred into a specific *lox* site in tobacco by biolistic bombardment and *Cre-lox* site specific recombination.

A construct or vector may include a plant promoter to express the protein or protein fragment of choice. A number of promoters which are active in plant cells have been described in the literature. These include the nopaline synthase (NOS) promoter (Ebert *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 84*:5745-5749 (1987), the entirety of which is herein incorporated by reference), the octopine synthase (OCS) promoter (which are carried on tumor-inducing plasmids of *Agrobacterium tumefaciens*), the caulimovirus promoters such as the cauliflower mosaic virus (CaMV) 19S promoter (Lawton *et al.*, *Plant Mol. Biol. 9*:315-324 (1987), the entirety of which is herein incorporated by reference) and the CAMV 35S promoter (Odell *et al.*, *Nature 313*:810-812 (1985) the entirety of which is herein incorporated by reference), the figwort mosaic virus 35S-promoter, the light-inducible promoter from the small subunit of ribulose-1,5-bis-phosphate carboxylase (ssRUBISCO), the Adh promoter (Walker *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 84*:6624-6628 (1987), the entirety of which is herein incorporated by reference), the sucrose synthase promoter (Yang *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 87*:4144-4148 (1990), the entirety of which is herein incorporated by reference), the R gene complex promoter (Chandler *et al.*, *The Plant Cell 1*:1175-1183 (1989), the entirety of which is herein incorporated by reference), and the chlorophyll a/b binding protein gene promoter, *etc.* These promoters have been used to create DNA constructs which have been expressed in plants; *see*, *e.g.*, PCT publication WO 84/02913, herein incorporated by reference in its entirety.

Promoters which are known or are found to cause transcription of DNA in plant cells can be used in the present invention. Such promoters may be obtained from a

variety of sources such as plants and plant viruses. It is preferred that the particular promoter selected should be capable of causing sufficient expression to result in the production of an effective amount of protein to cause the desired phenotype. In addition to promoters which are known to cause transcription of DNA in plant cells, other promoters may be identified for use in the current invention by screening a plant cDNA library for genes which are selectively or preferably expressed in the target tissues or cells.

For the purpose of expression in source tissues of the plant, such as the leaf, seed, root or stem, it is preferred that the promoters utilized in the present invention have relatively high expression in these specific tissues. For this purpose, one may choose from a number of promoters for genes with tissue- or cell-specific or -enhanced expression. Examples of such promoters reported in the literature include the chloroplast glutamine synthetase GS2 promoter from pea (Edwards *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 87*:3459-3463 (1990), herein incorporated by reference in its entirety), the chloroplast fructose-1,6-biphosphatase (FBPase) promoter from wheat (Lloyd *et al.*, *Mol. Gen. Genet. 225*:209-216 (1991), herein incorporated by reference in its entirety), the nuclear photosynthetic ST-LS1 promoter from potato (Stockhaus *et al.*, *EMBO J. 8*:2445-2451 (1989), herein incorporated by reference in its entirety), the phenylalanine ammonia-lyase (PAL) promoter and the chalcone synthase (CHS) promoter from *Arabidopsis thaliana*. Also reported to be active in photosynthetically active tissues are the ribulose-1,5-bisphosphate carboxylase (RbcS) promoter from eastern larch (*Larix laricina*), the promoter for the *cab* gene, cab6, from pine (Yamamoto *et al.*, *Plant Cell Physiol. 35*:773-778 (1994), herein incorporated by reference in its entirety), the promoter for the Cab-1 gene from wheat (Fejes *et al.*, *Plant Mol. Biol. 15*:921-932 (1990), herein incorporated by reference in its entirety), the promoter for the CAB-1 gene from spinach (Lubberstedt *et al.*, *Plant Physiol. 104*:997-1006 (1994), herein incorporated by reference in its entirety), the promoter for the cab1R gene from rice (Luan *et al.*, *Plant Cell. 4*:971-

981 (1992), the entirety of which is herein incorporated by reference), the pyruvate, orthophosphate dikinase (PPDK) promoter from maize (Matsuoka *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 90*:9586-9590 (1993), herein incorporated by reference in its entirety), the promoter for the tobacco Lhcb1*2 gene (Cerdan *et al.*, *Plant Mol. Biol. 33*:245-255. (1997), herein incorporated by reference in its entirety), the *Arabidopsis thaliana* SUC2 sucrose-H+ symporter promoter (Truernit *et al.*, *Planta. 196*:564-570 (1995), herein incorporated by reference in its entirety), and the promoter for the thylacoid membrane proteins from spinach (psaD, psaF, psaE, PC, FNR, atpC, atpD, cab, rbcS). Other promoters for the chlorophyll a/b-binding proteins may also be utilized in the present invention, such as the promoters for LhcB gene and PsbP gene from white mustard (*Sinapis alba*; Kretsch *et al.*, *Plant Mol. Biol. 28*:219-229 (1995), the entirety of which is herein incorporated by reference).

For the purpose of expression in sink tissues of the plant, such as the tuber of the potato plant, the fruit of tomato, or the seed of maize, wheat, rice, and barley, it is preferred that the promoters utilized in the present invention have relatively high expression in these specific tissues. A number of promoters for genes with tuber-specific or -enhanced expression are known, including the class I patatin promoter (Bevan *et al.*, *EMBO J. 8*:1899-1906 (1986); Jefferson *et al.*, *Plant Mol. Biol. 14*995-1006 (1990), both of which are herein incorporated by reference in its entirety), the promoter for the potato tuber ADPGPP genes, both the large and small subunits, the sucrose synthase promoter (Salanoubat and Belliard, *Gene. 60*:47-56 (1987), Salanoubat and Belliard, *Gene. 84*:181-185 (1989), both of which are incorporated by reference in their entirety), the promoter for the major tuber proteins including the 22 kd protein complexes and proteinase inhibitors (Hannapel, *Plant Physiol. 101*:703-704 (1993), herein incorporated by reference in its entirety), the promoter for the granule bound starch synthase gene (GBSS) (Visser *et al.*, *Plant Mol. Biol. 17*:691-699 (1991), herein incorporated by reference in its entirety), and other class I and II patatins promoters (Koster-Topfer *et al.*, *Mol. Gen.*

Genet. 219:390-396 (1989); Mignery *et al.*, *Gene. 62*:27-44 (1988), both of which are herein incorporated by reference in their entirety).

Other promoters can also be used to express a fructose 1,6 bisphosphate aldolase gene in specific tissues, such as seeds or fruits. The promoter for β-conglycinin (Chen *et al.*, *Dev. Genet. 10*:112-122 (1989), herein incorporated by reference in its entirety) or other seed-specific promoters such as the napin and phaseolin promoters, can be used. The zeins are a group of storage proteins found in maize endosperm. Genomic clones for zein genes have been isolated (Pedersen *et al.*, *Cell 29*:1015-1026 (1982), herein incorporated by reference in its entirety), and the promoters from these clones, including the 15 kD, 16 kD, 19 kD, 22 kD, 27 kD, and gamma genes, could also be used. Other promoters known to function, for example, in maize, include the promoters for the following genes: *waxy, Brittle, Shrunken 2*, Branching enzymes I and II, starch synthases, debranching enzymes, oleosins, glutelins, and sucrose synthases. A particularly preferred promoter for maize endosperm expression is the promoter for the glutelin gene from rice, more particularly the Osgt-1 promoter (Zheng *et al.*, *Mol. Cell Biol. 13*:5829-5842 (1993), herein incorporated by reference in its entirety). Examples of promoters suitable for expression in wheat include those promoters for the ADPglucose pyrophosphorylase (ADPGPP) subunits, the granule bound and other starch synthases, the branching and debranching enzymes, the embryogenesis-abundant proteins, the gliadins, and the glutenins. Examples of such promoters in rice include those promoters for the ADPGPP subunits, the granule bound and other starch synthases, the branching enzymes, the debranching enzymes, sucrose synthases, and the glutelins. A particularly preferred promoter is the promoter for rice glutelin, Osgt-1. Examples of such promoters for barley include those for the ADPGPP subunits, the granule bound and other starch synthases, the branching enzymes, the debranching enzymes, sucrose synthases, the hordeins, the embryo globulins, and the aleurone specific proteins.

Root specific promoters may also be used. An example of such a promoter is the promoter for the acid chitinase gene (Samac *et al.*, *Plant Mol. Biol. 25*:587-596 (1994), the entirety of which is herein incorporated by reference). Expression in root tissue could also be accomplished by utilizing the root specific subdomains of the CaMV35S promoter that have been identified (Lam *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.) 86*:7890-7894 (1989), herein incorporated by reference in its entirety). Other root cell specific promoters include those reported by Conkling *et al.* (Conkling *et al.*, *Plant Physiol. 93*:1203-1211 (1990), the entirety of which is herein incorporated by reference).

Additional promoters that may be utilized are described, for example, in U.S. Patent Nos. 5,378,619, 5,391,725, 5,428,147, 5,447,858, 5,608,144, 5,608,144, 5,614,399, 5,633,441, 5,633,435, and 4,633,436, all of which are herein incorporated in their entirety. In addition, a tissue specific enhancer may be used (Fromm *et al.*, *The Plant Cell 1*:977-984 (1989), the entirety of which is herein incorporated by reference). Further promoters are described in Agents of the Invention Section (a)(i). It is further understood that one or more of the promoters of the present invention may be used.

Constructs or vectors may also include, with the coding region of interest, a nucleic acid sequence that acts, in whole or in part, to terminate transcription of that region. For example, such sequences have been isolated including the Tr7 3' sequence and the nos 3' sequence (Ingelbrecht *et al.*, *The Plant Cell 1*:671-680 (1989), the entirety of which is herein incorporated by reference; Bevan *et al.*, *Nucleic Acids Res.. 11*:369-385 (983), the entirety of which is herein incorporated by reference), or the like. It is understood that one or more sequences of the present invention that act, to terminate transcription may be used.

A vector or construct may also include other regulatory elements. Examples of such include the Adh intron 1 (Callis *et al.*, *Genes and Develop. 1*:1183-1200 (1987), the entirety of which is herein incorporated by reference), the sucrose synthase intron (Vasil *et al.*, Plant Physiol. 91:1575-1579 (1989), the entirety of which is herein incorporated by

reference) and the TMV omega element (Gallie *et al.*, *The Plant Cell 1*:301-311 (1989), the entirety of which is herein incorporated by reference). These and other regulatory elements may be included when appropriate. It is also understood that one or more of the regulatory regions of the present invention may be used.

A vector or construct may also include a selectable marker. Selectable markers may also be used to select for plants or plant cells that contain the exogenous genetic material. Examples of such include, but are not limited to, a neo gene (Potrykus *et al.*, *Mol. Gen. Genet. 199*:183-188 (1985), the entirety of which is herein incorporated by reference) which codes for kanamycin resistance and can be selected for using kanamycin, G418, *etc.*; a bar gene which codes for bialaphos resistance; a mutant EPSP synthase gene (Hinchee *et al.*, *Bio/Technology 6*:915-922 (1988), the entirety of which is herein incorporated by reference) which encodes glyphosate resistance; a nitrilase gene which confers resistance to bromoxynil (Stalker *et al.*, *J. Biol. Chem. 263*:6310-6314 (1988), the entirety of which is herein incorporated by reference); a mutant acetolactate synthase gene (ALS) which confers imidazolinone or sulphonylurea resistance (European Patent Application 154,204 (Sept. 11, 1985), the entirety of which is herein incorporated by reference); and a methotrexate resistant DHFR gene (Thillet *et al.*, *J. Biol. Chem. 263*:12500-12508 (1988), the entirety of which is herein incorporated by reference).

A vector or construct may also include a transit peptide. Incorporation of a suitable chloroplast transit peptide may also be employed (European Patent Application Publication Number 0218571, the entirety of which is herein incorporated by reference). Translational enhancers may also be incorporated as part of the vector DNA. DNA constructs could contain one or more 5' non-translated leader sequences which may serve to enhance expression of the gene products from the resulting mRNA transcripts. Such sequences may be derived from the promoter selected to express the gene or can be specifically modified to increase translation of the mRNA. Such regions may also be obtained from viral RNAs, from suitable eukaryotic genes, or from a synthetic gene

sequence. For a review of optimizing expression of transgenes, *see* Koziel *et al., Plant Mol. Biol. 32*:393-405 (1996), the entirety of which is herein incorporated by reference.

A vector or construct may also include a screenable marker. Screenable markers may be used to monitor expression. Exemplary screenable markers include a β-glucuronidase or uidA gene (GUS) which encodes an enzyme for which various chromogenic substrates are known (Jefferson, *Plant Mol. Biol, Rep. 5*:387-405 (1987), the entirety of which is herein incorporated by reference; Jefferson *et al., EMBO J. 6*:3901-3907 (1987), the entirety of which is herein incorporated by reference); an R-locus gene, which encodes a product that regulates the production of anthocyanin pigments (red color) in plant tissues ((Dellaporta *et al.*, Stadler Symposium *11*:263-282 (1988), the entirety of which is herein incorporated by reference); a β-lactamase gene (Sutcliffe *et al., Proc. Natl. Acad. Sci. (U.S.A.) 75*:3737-3741 (1978), the entirety of which is herein incorporated by reference), a gene which encodes an enzyme for which various chromogenic substrates are known (*e.g.*, PADAC, a chromogenic cephalosporin); a luciferase gene (Ow *et al., Science 234*:856-859 (1986), the entirety of which is herein incorporated by reference) a xylE gene (Zukowsky *et al., Proc. Natl. Acad. Sci. U.S.A. 80*:1101-1105 (1983), the entirety of which is herein incorporated by reference) which encodes a catechol dioxygenase that can convert chromogenic catechols; an α-amylase gene (Ikatu *et al., Bio/Technol. 8*:241-242 (1990), the entirety of which is herein incorporated by reference); a tyrosinase gene (Katz *et al., J. Gen. Microbiol. 129*:2703-2714 (1983), the entirety of which is herein incorporated by reference) which encodes an enzyme capable of oxidizing tyrosine to DOPA and dopaquinone which in turn condenses to melanin; an α-galactosidase, which will turn a chromogenic α-galactose substrate.

Included within the terms "selectable or screenable marker genes" are also genes which encode a secretable marker whose secretion can be detected as a means of identifying or selecting for transformed cells. Examples include markers which encode a secretable antigen that can be identified by antibody interaction, or even secretable

enzymes which can be detected catalytically. Secretable proteins fall into a number of classes, including small, diffusible proteins detectable, *e.g.*, by ELISA, small active enzymes detectable in extracellular solution (*e.g.*, α-amylase, β-lactamase, phosphinothricin transferase), or proteins which are inserted or trapped in the cell wall (such as proteins which include a leader sequence such as that found in the expression unit of extension or tobacco PR-S). Other possible selectable and/or screenable marker genes will be apparent to those of skill in the art.

Methods and compositions for transforming a bacteria and other microorganisms are known in the art (see for example Sambrook *et al., Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., (1989), the entirety of which is herein incorporated by reference).

There are many methods for introducing transforming nucleic acid molecules into plant cells. Suitable methods are believed to include virtually any method by which nucleic acid molecules may be introduced into a cell, such as by *Agrobacterium* infection or direct delivery of nucleic acid molecules such as, for example, by PEG-mediated transformation, by electroporation or by acceleration of DNA coated particles, *etc.* (Potrykus, *Ann. Rev. Plant Physiol. Plant Mol. Biol. 42*:205-225 (1991), the entirety of which is herein incorporated by reference; Vasil, *Plant Mol. Biol. 25*:925-937 (1994), the entirety of which is herein incorporated by reference. For example, electroporation has been used to transform maize protoplasts (Fromm *et al.*, *Nature 312*:791-793 (1986), the entirety of which is herein incorporated by reference).

Technology for introduction of DNA into cells is well known to those of skill in the art. Four general methods for delivering a gene into cells have been described: (1) chemical methods (Graham and van der Eb, *Virology, 54*:536-539 (1973), the entirety of which is herein incorporated by reference); (2) physical methods such as microinjection (Capecchi, *Cell 22*:479-488 (1980), electroporation (Wong and Neumann, *Biochem. Biophys. Res. Commun., 107*:584-587 (1982); Fromm *et al., Proc. Natl. Acad. Sci.*

U.S.A., 82:5824-5828 (1985); U. S. Patent No. 5,384,253; and the gene gun (Johnston and Tang, *Methods Cell Biol. 43*:353-365 (1994), all of which the entirety is herein incorporated by reference; (3) viral vectors (Clapp, *Clin. Perinatol., 20*:155-168 (1993); Lu *et al., J. Exp. Med., 178*:2089-2096 (1993); Eglitis and Anderson, *Biotechniques, 6*:608-614 (1988), all of which the entirety is herein incorporated by reference); and (4) receptor-mediated mechanisms (Curiel *et al., Hum. Gen. Ther., 3*:147-154 (1992); Wagner *et al., Proc. Natl. Acad. Sci. (U.S.A.) 89*:6099-6103 (1992), all of which the entirety is herein incorporated by reference).

Acceleration methods that may be used include, for example, microprojectile bombardment and the like. One example of a method for delivering transforming nucleic acid molecules to plant cells is microprojectile bombardment. This method has been reviewed by Yang and Christou, eds., *Particle Bombardment Technology for Gene Transfer*, Oxford Press, Oxford, England (1994), the entirety of which is herein incorporated by reference). Non-biological particles (microprojectiles) that may be coated with nucleic acids and delivered into cells by a propelling force. Exemplary particles include those comprised of tungsten, gold, platinum, and the like.

A particular advantage of microprojectile bombardment, in addition to it being an effective means of reproducibly, and stably transforming monocotyledons, is that neither the isolation of protoplasts (Cristou *et al., Plant Physiol. 87*:671-674 (1988), the entirety of which is herein incorporated by reference) nor the susceptibility of *Agrobacterium* infection is required. An illustrative embodiment of a method for delivering DNA into maize cells by acceleration is a biolistics γ-particle delivery system, which can be used to propel particles coated with DNA through a screen, such as a stainless steel or Nytex screen, onto a filter surface covered with corn cells cultured in suspension. Gordon-Kamm *et al.*, describes the basic procedure for coating tungsten particles with DNA (Gordon-Kamm *et al., Plant Cell 2*:603-618 (1990), the entirety of which is herein incorporated by reference). The screen disperses the tungsten nucleic acid particles so

that they are not delivered to the recipient cells in large aggregates. A particle delivery system suitable for use with the present invention is the helium acceleration PDS-1000/He gun which is available from Bio-Rad Laboratories (Bio-Rad, Hercules, California)(Sanford *et al.*, *Technique 3*:3-16 (1991), the entirety of which is herein incorporated by reference).

For the bombardment, cells in suspension may be concentrated on filters. Filters containing the cells to be bombarded are positioned at an appropriate distance below the microprojectile stopping plate. If desired, one or more screens are also positioned between the gun and the cells to be bombarded.

Alternatively, immature embryos or other target cells may be arranged on solid culture medium. The cells to be bombarded are positioned at an appropriate distance below the macroprojectile stopping plate. If desired, one or more screens are also positioned between the acceleration device and the cells to be bombarded. Through the use of techniques set forth herein one may obtain up to 1000 or more foci of cells transiently expressing a marker gene. The number of cells in a focus which express the exogenous gene product 48 hours post-bombardment often range from one to ten and average one to three.

In bombardment transformation, one may optimize the prebombardment culturing conditions and the bombardment parameters to yield the maximum numbers of stable transformants. Both the physical and biological parameters for bombardment are important in this technology. Physical factors are those that involve manipulating the DNA/microprojectile precipitate or those that affect the flight and velocity of either the macro- or microprojectiles. Biological factors include all steps involved in manipulation of cells before and immediately after bombardment, the osmotic adjustment of target cells to help alleviate the trauma associated with bombardment, and also the nature of the transforming DNA, such as linearized DNA or intact supercoiled plasmids. It is believed

that pre-bombardment manipulations are especially important for successful transformation of immature embryos.

In another alternative embodiment, plastids can be stably transformed. Methods disclosed for plastid transformation in higher plants include the particle gun delivery of DNA containing a selectable marker and targeting of the DNA to the plastid genome through homologous recombination (Svab *et al. Proc. Natl. Acad. Sci. (U.S.A.) 87*:8526-8530 (1990): Svab and Maliga *Proc. Natl. Acad. Sci. (U.S.A.) 90*:913-917 (1993)); (Staub, J. M. and Maliga, P. *EMBO J. 12*:601-606 (1993), U.S. Patents 5, 451,513 and 5,545,818, all of which are herein incorporated by reference in their entirety).

Accordingly, it is contemplated that one may wish to adjust various aspects of the bombardment parameters in small scale studies to fully optimize the conditions. One may particularly wish to adjust physical parameters such as gap distance, flight distance, tissue distance, and helium pressure. One may also minimize the trauma reduction factors by modifying conditions which influence the physiological state of the recipient cells and which may therefore influence transformation and integration efficiencies. For example, the osmotic state, tissue hydration and the subculture stage or cell cycle of the recipient cells may be adjusted for optimum transformation. The execution of other routine adjustments will be known to those of skill in the art in light of the present disclosure.

Agrobacterium-mediated transfer is a widely applicable system for introducing genes into plant cells because the DNA can be introduced into whole plant tissues, thereby bypassing the need for regeneration of an intact plant from a protoplast. The use of *Agrobacterium*-mediated plant integrating vectors to introduce DNA into plant cells is well known in the art. See, for example the methods described (Fraley *et al.*, *Biotechnology 3*:629-635 (1985); Rogers *et al., Meth. Enzymol. 153*:253-277 (1987), both of which are herein incorporated by reference in their entirety. Further, the integration of the Ti-DNA is a relatively precise process resulting in few rearrangements. The region of

DNA to be transferred is defined by the border sequences, and intervening DNA is usually inserted into the plant genome as described (Spielmann *et al., Mol. Gen. Genet., 205*:34 (1986), the entirety of which is herein incorporated by reference).

Modern *Agrobacterium* transformation vectors are capable of replication in *E. coli* as well as *Agrobacterium*, allowing for convenient manipulations as described (Klee *et al., In: Plant DNA Infectious Agents*, T. Hohn and J. Schell, eds., Springer-Verlag, New York, pp. 179-203 (1985), the entirety of which is herein incorporated by reference). Moreover, recent technological advances in vectors for *Agrobacterium*-mediated gene transfer have improved the arrangement of genes and restriction sites in the vectors to facilitate construction of vectors capable of expressing various polypeptide coding genes. The vectors described have convenient multi-linker regions flanked by a promoter and a polyadenylation site for direct expression of inserted polypeptide coding genes and are suitable for present purposes (Rogers *et al., Meth. Enzymo. 153*:253-277 (1987), the entirety of which is herein incorporated by reference). In addition, *Agrobacterium* containing both armed and disarmed Ti genes can be used for the transformations. In those plant strains where *Agrobacterium*-mediated transformation is efficient, it is the method of choice because of the facile and defined nature of the gene transfer.

A transgenic plant formed using *Agrobacterium* transformation methods typically contains a single gene on one chromosome. Such transgenic plants can be referred to as being heterozygous for the added gene. More preferred is a transgenic plant that is homozygous for the added structural gene; *i.e.*, a transgenic plant that contains two added genes, one gene at the same locus on each chromosome of a chromosome pair. A homozygous transgenic plant can be obtained by sexually mating (selfing) an independent segregant transgenic plant that contains a single added gene, germinating some of the seed produced and analyzing the resulting plants produced for the gene of interest.

It is also to be understood that two different transgenic plants can also be mated to produce offspring that contain two independently segregating added, exogenous genes.

Selfing of appropriate progeny can produce plants that are homozygous for both added, exogenous genes that encode a polypeptide of interest. Back-crossing to a parental plant and out-crossing with a non-transgenic plant are also contemplated, as is vegetative propagation.

Transformation of plant protoplasts can be achieved using methods based on calcium phosphate precipitation, polyethylene glycol treatment, electroporation, and combinations of these treatments. See for example, Potrykus *et al. Mol. Gen. Genet. 205*:193-200 (1986); Lorz *et al., Mol. Gen. Genet. 199*:178 (1985); Fromm *et al., Nature 319*:791 (1986); Uchimiya *et al., Mol. Gen. Genet. 204*:204 (1986); Callis *et al., Genes and Development 1183*: (1987); and Marcotte *et al., Nature 335*:454 (1988), all of which the entirety is herein incorporated by reference.

Application of these systems to different plant strains depends upon the ability to regenerate that particular plant strain from protoplasts. Illustrative methods for the regeneration of cereals from protoplasts are described (Fujimura *et al., Plant Tissue Culture Letters 2*:74 (1985); Toriyama *et al., Theor Appl. Genet. 205*:34 (1986); Yamada *et al., Plant Cell Rep. 4*:85 (1986); Abdullah *et al., Biotechnology 4*:1087 (1986), all of which the entirety is herein incorporated by reference).

To transform plant strains that cannot be successfully regenerated from protoplasts, other ways to introduce DNA into intact cells or tissues can be utilized. For example, regeneration of cereals from immature embryos or explants can be effected as described (Vasil, *Biotechnology 6*:397 (1988), the entirety of which is herein incorporated by reference). In addition, "particle gun" or high-velocity microprojectile technology can be utilized (Vasil *et al., Bio/Technology 10*:667 (1992), the entirety of which is herein incorporated by reference).

Using the latter technology, DNA is carried through the cell wall and into the cytoplasm on the surface of small metal particles as described (Klein *et al., Nature, 328*:70, (1987); Klein *et al., Proc. Natl. Acad. Sci. (U.S.A.)* 85:8502-8505 (1988);

McCabe *et al.*, *Biotechnolog, 6*:923 (1988), all of which the entirety is herein incorporated by reference). The metal particles penetrate through several layers of cells and thus allow the transformation of cells within tissue explants.

Other methods of cell transformation can also be used and include but are not limited to introduction of DNA into plants by direct DNA transfer into pollen (Hess *et al.*, *Intern Rev. Cytol, 107*:367 (1987); Luo *et al.*, *Plant Mol. Biol. Reporter 6*:165 (1988), all of which the entirety is herein incorporated by reference), by direct injection of DNA into reproductive organs of a plant (Pena *et al.*, *Nature 325*:274 (1987), the entirety of which is herein incorporated by reference), or by direct injection of DNA into the cells of immature embryos followed by the rehydration of dessicated embryos (Neuhaus *et al.*, *Theor. Appl. Genet. 75*:30 (1987), the entirety of which is herein incorporated by reference).

The regeneration, development, and cultivation of plants from single plant protoplast transformants or from various transformed explants is well known in the art (Weissbach and Weissbach, *In: Methods for Plant Molecular Biology*, (Eds.), Academic Press, Inc. San Diego, CA, (1988), the entirety of which is herein incorporated by reference). This regeneration and growth process typically includes the steps of selection of transformed cells, culturing those individualized cells through the usual stages of embryonic development through the rooted plantlet stage. Transgenic embryos and seeds are similarly regenerated. The resulting transgenic rooted shoots are thereafter planted in an appropriate plant growth medium such as soil.

The development or regeneration of plants containing the foreign, exogenous gene that encodes a protein of interest is well known in the art. Preferably, the regenerated plants are self-pollinated to provide homozygous transgenic plants, as discussed before. Otherwise, pollen obtained from the regenerated plants is crossed to seed-grown plants of agronomically important lines. Conversely, pollen from plants of these important lines is used to pollinate regenerated plants. A transgenic plant of the present invention

containing a desired polypeptide is cultivated using methods well known to one skilled in the art.

There are a variety of methods for the regeneration of plants from plant tissue. The particular method of regeneration will depend on the starting plant tissue and the particular plant species to be regenerated.

Methods for transforming dicots, primarily by use of *Agrobacterium tumefaciens*, and obtaining transgenic plants have been published for cotton (U. S. Patent No. 5,004,863, U.S. Patent No. 5,159,135, U.S. Patent No. 5,518,908, all of which the entirety is herein incorporated by reference); soybean (U. S. Patent No. 5,569,834, U. S. Patent No. 5,416,011, McCabe *et al., Biotechnology 6*:923 (1988), Christou *et al., Plant Physiol. 87*:671-674 (1988), all of which the entirety is herein incorporated by reference); *Brassica* (U. S. Patent No. 5,463,174, the entirety of which is herein incorporated by reference); peanut (Cheng *et al., Plant Cell Rep. 15*:653-657 (1996), McKently *et al., Plant Cell Rep. 14*:699-703 (1995), all of which the entirety is herein incorporated by reference); papaya (Yang *et al.*, (1996), the entirety of which is herein incorporated by reference); pea (Grant *et al., Plant Cell Rep. 15*:254-258, (1995), the entirety of which is herein incorporated by reference).

Transformation of monocotyledons using electroporation, particle bombardment, and *Agrobacterium* have also been reported. Transformation and plant regeneration have been achieved in asparagus (Bytebier *et al., Proc. Natl. Acad. Sci. (U.S.A.) 84*:5345 (1987), the entirety of which is herein incorporated by reference); barley (Wan and Lemaux, *Plant Physiol 104*:37 (1994), the entirety of which is herein incorporated by reference); maize (Rhodes *et al., Science 240*:204 (1988), Gordon-Kamm *et al., Plant Cell, 2*:603 (1990), Fromm *et al., Bio/Technology 8*:833 (1990), Koziel *et al., Bio/Technology 11*:194 (1993), Armstrong *et al., Crop Science 35*:550-557 (1995), all of which the entirety is herein incorporated by reference); oat (Somers *et al., Bio/Technology, 10*:1589 (1992), the entirety of which is herein incorporated by

reference); orchardgrass (Horn *et al., Plant Cell Rep. 7*:469, (1988), the entirety of which is herein incorporated by reference); rice (Toriyama *et al., Theor Appl. Genet. 205*:34 (1986); Park *et al., Plant Mol. Biol. 32*:1135-1148 (1996); Abedinia *et al., Aust. J. Plant Physiol.24*:133-141 (1997); Zhang and Wu, *Theor. Appl. Genet. 76*:835, (1988); Zhang *et al. Plant Cell Rep. 7*:379, (1988); Battraw and Hall, *Plant Sci. 86*:191-202 (1992); Christou *et al., Bio/Technology 9*:957 (1991), all of which the entirety is herein incorporated by reference); sugarcane (Bower and Birch, *Plant J. 2*:409 (1992), the entirety of which is herein incorporated by reference); tall fescue (Wang *et al., Bio/Technology 10*:691 (1992), the entirety of which is herein incorporated by reference), and wheat (Vasil *et al., Bio/Technology 10*:667 (1992), the entirety of which is herein incorporated by reference; U. S. Patent No. 5,631,152, the entirety of which is herein incorporated by reference).

Assays for gene expression based on the transient expression of cloned nucleic acid constructs have been developed by introducing the nucleic acid molecules into plant cells by polyethylene glycol treatment, electroporation, or particle bombardment (Marcotte *et al., Nature 335*:454-457 (1988), the entirety of which is herein incorporated by reference; Marcotte *et al., Plant Cell 1*:523-532 (1989), the entirety of which is herein incorporated by reference; McCarty *et al., Cell 66*:895-905 (1991), the entirety of which is herein incorporated by reference; Hattori *et al., Genes Dev. 6*:609-618 (1992), the entirety of which is herein incorporated by reference; Goff *et al., EMBO J. 9*:2517-2522 (1990), the entirety of which is herein incorporated by reference). Transient expression systems may be used to functionally dissect gene constructs (*See generally*, Mailga *et al., Methods in Plant Molecular Biology,* Cold Spring Harbor Press (1995)).

Any of the nucleic acid molecules of the present invention may be introduced into a plant cell in a permanent or transient manner in combination with other genetic elements such as vectors, promoters enhancers *etc.* Further any of the nucleic acid molecules of the present invention may be introduced into a plant cell in a manner that

allows for over expression of the protein or fragment thereof encoded by the nucleic acid molecule.

Nucleic acid molecules of the present invention may be used in cosuppression. Cosuppression is the reduction in expression levels, usually at the level of RNA, of a particular endogenous gene or gene family by the expression of a homologous sense construct that is capable of transcribing mRNA of the same strandedness as the transcript of the endogenous gene (Napoli *et al., Plant Cell* 2:279-289 (1990), the entirety of which is herein incorporated by reference; van der Krol *et al., Plant Cell* 2:291-299 (1990), the entirety of which is herein incorporated by reference). Cosuppression may result from stable transformation with a single copy nucleic acid molecule that is homologous to a nucleic acid sequence found with the cell (Prolls and Meyer, *Plant J.*. 2:465-475 (1992), the entirety of which is herein incorporated by reference) or with multiple copies of a nucleic acid molecule that is homologous to a nucleic acid sequence found with the cell (Mittlesten *et al., Mol. Gen. Genet. 244*:325-330 (1994), the entirety of which is herein incorporated by reference). Genes, even though different, linked to homologous promoters may result in the cosuppression of the linked genes (Vaucheret, *C.R. Acad. Sci. III 316*: 1471-1483 (1993), the entirety of which is herein incorporated by reference).

This technique has, for example been applied to generate white flowers from red petunia and tomatoes that do not ripen on the vine. Up to 50% of petunia transformants that contained a sense copy of the chalcone synthase (CHS) gene produced white flowers or floral sectors; this was as a result of the post-transcriptional loss of mRNA encoding CHS (Flavell, *Proc. Natl. Acad. Sci. (U.S.A.) 91*:3490-3496 (1994)), the entirety of which is herein incorporated by reference). Cosuppression may require the coordinate transcription of the transgene and the endogenous gene, and can be reset by a developmental control mechanism (Jorgensen, *Trends Biotechnol. 8*:340344 (1990), the entirety of which is herein incorporated by reference; Meins and Kunz, In: *Gene Inactivation and Homologous Recombination in Plants* (Paszkowski, J., ed.), pp. 335-

348. Kluwer Academic, Netherlands (1994), the entirety of which is herein incorporated by reference).

It is understood that one or more of the nucleic acids of the present invention comprising an protein coding region or fragment thereof located within SEQ ID NO: 1 through SEQ ID NO: 51,470 or complement thereof or fragment of either, may be introduced into a plant cell and transcribed using an appropriate promoter with such transcription resulting in the co-suppression of an endogenous protein.

Nucleic acid molecules of the present invention may be used to reduce gene function. Antisense approaches are a way of preventing or reducing gene function by targeting the genetic material (Mol *et al., FEBS Lett. 268*:427-430 (1990), the entirety of which is herein incorporated by reference). The objective of the antisense approach is to use a sequence complementary to the target gene to block its expression and create a mutant cell line or organism in which the level of a single chosen protein is selectively reduced or abolished. Antisense techniques have several advantages over other 'reverse genetic' approaches. The site of inactivation and its developmental effect can be manipulated by the choice of promoter for antisense genes or by the timing of external application or microinjection. Antisense can manipulate its specificity by selecting either unique regions of the target gene or regions where it shares homology to other related genes (Hiatt *et al., In Genetic Engineering,* Setlow (ed.), Vol. 11, New York: Plenum 49-63 (1989), the entirety of which is herein incorporated by reference).

The principle of regulation by antisense RNA is that RNA that is complementary to the target mRNA is introduced into cells, resulting in specific RNA:RNA duplexes being formed by base pairing between the antisense substrate and the target mRNA (Green *et al., Annu. Rev. Biochem. 55*:569-597 (1986), the entirety of which is herein incorporated by reference). Under one embodiment, the process involves the introduction and expression of an antisense gene sequence. Such a sequence is one in which part or all of the normal gene sequences are placed under a promoter in inverted orientation so that

the 'wrong' or complementary strand is transcribed into a noncoding antisense RNA that hybridizes with the target mRNA and interferes with its expression (Takayama and Inouye, *Crit. Rev. Biochem. Mol. Biol. 25*:155-184 (1990), the entirety of which is herein incorporated by reference). An antisense vector is constructed by standard procedures and introduced into cells by transformation, transfection, electroporation, microinjection, or by infection, *etc.* The type of transformation and choice of vector will determine whether expression is transient or stable. The promoter used for the antisense gene may influence the level, timing, tissue, specificity, or inducibility of the antisense inhibition.

It is understood that protein synthesis activity in a plant cell may be reduced or depressed by growing a transformed plant cell containing a nucleic acid molecule of the present invention.

Antibodies have been expressed in plants (Hiatt *et al., Nature 342:*76-78 (1989), the entirety of which is herein incorporated by reference; Conrad and Fielder, *Plant Mol. Biol. 26:*1023-1030 (1994), the entirety of which is herein incorporated by reference). Cytoplasmic expression of a scFv (single-chain Fv antibodies) has been reported to delay infection by artichoke mottled crinkle virus. Transgenic plants that express antibodies directed against endogenous proteins may exhibit a physiological effect (Philips *et al., EMBO J. 16:*4489-4496 (1997), the entirety of which is herein incorporated by reference; Marion-Poll, *Trends in Plant Science 2*:447-448 (1997), the entirety of which is herein incorporated by reference). For example, expressed anti-abscisic antibodies reportedly result in a general perturbation of seed development (Philips *et al., EMBO J. 16:*4489-4496 (1997)).

Antibodies that are catalytic may also be expressed in plants (abzymes). The principle behind abzymes is that since antibodies may be raised against many molecules, this recognition ability can be directed toward generating antibodies that bind transition states to force a chemical reaction forward (Persidas, *Nature Biotechnology 15*:1313-1315 (1997), the entirety of which is herein incorporated by reference; Baca *et al., Ann.*

Rev. Biophys. Biomol. Struct. 26:461-493 (1997), the entirety of which is herein incorporated by reference). The catalytic abilities of abzymes may be enhanced by site directed mutagenesis. Examples of abzymes are, for example, set forth in U.S. Patent No: 5,658,753; U.S. Patent No. 5,632,990; U.S. Patent No. 5,631,137; U.S. Patent 5,602,015; U.S. Patent No. 5,559,538; U.S. Patent No. 5,576,174; U.S. Patent No. 5,500,358; U.S. Patent 5,318,897; U.S. Patent No. 5,298,409; U.S. Patent No. 5,258,289 and U.S. Patent No. 5,194,585, all of which are herein incorporated in their entirety.

It is understood that any of the antibodies of the present invention may be expressed in plants and that such expression can result in a physiological effect. It is also understood that any of the expressed antibodies may be catalytic.

In addition to the above discussed procedures, practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of macromolecules (*e.g.*, DNA molecules, plasmids, *etc.*), generation of recombinant organisms and the screening and isolating of clones, (see for example, Sambrook *et al., Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Press (1989); Mailga *et al., Methods in Plant Molecular Biology,* Cold Spring Harbor Press (1995), the entirety of which is herein incorporated by reference; Birren *et al., Genome Analysis: Analyzing DNA,* 1, Cold Spring Harbor, New York, the entirety of which is herein incorporated by reference).

The nucleotide sequence provided in SEQ ID NO: 1, through SEQ ID NO: 51,470 or fragment thereof, or complement thereof, or a nucleotide sequence at least 90% identical, preferably 95%, identical even more preferably 99% or 100% identical to the sequence provided in SEQ ID NO: 1 through SEQ ID NO: 51,470 or fragment thereof, or complement thereof, can be "provided" in a variety of mediums to facilitate use. Such a medium can also provide a subset thereof in a form that allows a skilled artisan to examine the sequences.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc, storage medium, and magnetic tape: optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate media comprising the nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word,, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (*e.g.*, text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing one or more of nucleotide sequences of the present invention, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al., J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al., Comp. Chem. 17*:203-207 (1993), the entirety of which is herein incorporated by reference) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs or proteins from other organisms. Such ORFs are protein-encoding fragments within the sequences of the present invention and are useful in producing commercially important proteins such as enzymes used in amino acid biosynthesis, metabolism, transcription, translation, RNA processing, nucleic acid and a protein degradation, protein modification, and DNA replication, restriction, modification, recombination, and repair.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the nucleic acid molecule of the present invention. As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As indicated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing

a search means. As used herein, "data storage means" refers to memory that can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention. As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the sequence of the present invention that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTIN and BLASTIX (NCBIA). One of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that during searches for commercially important fragments of the nucleic acid molecules of the present invention, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequences the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymatic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, *cis* elements, hairpin structures and inducible expression elements (protein binding sequences).

Thus, the present invention further provides an input means for receiving a target sequence, a data storage means for storing the target sequences of the present invention sequence identified using a search means as described above, and an output means for outputting the identified homologous sequences. A variety of structural formats for the input and output means can be used to input and output information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the sequence of the present invention by varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments sequence of the present invention. For example, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al., J. Mol. Biol. 215:*403-410 (1990)) can be used to identify open frames within the nucleic acid molecules of the present invention. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

Having now generally described the invention, the same will be more readily understood through reference to the following examples which are provided by way of illustration, and are not intended to be limiting of the present invention, unless specified.

Example 1

Genomic DNA Library

DNA Preparation

DNA from *Arabidopsis thaliana, Columbia* seedlings is prepared by a CTAB genomic DNA isolation protocol as described by Dean *et al. Plant J* 2:69-81(1992) and modified by Dubois *et al. Plant J. 13*:141-151 (1998).

A solution of DNA to be sheared is prepared in a 1.5 ml microcentrifuge tube by mixing 15 ug of DNA, 6 µl of 10X mung bean (MB) buffer (10X MB buffer = 300mM NaOAc, pH 5.0, 500 mM NaCl, 10 mM $ZnCl_2$, 50% glycerol), and water to a final volume of 60 µl. The DNA solution is kept on ice prior to sonication. For sonication, a cup horn probe chilled with ice water for 1 hour prior to sonication is used. The sonicator (Ultrasonic Liquid Processor XL2020 , Misonix Inc.) is pulsed for approximately 10 seconds on full power prior to use. DNA samples are sonicated twice for 6 seconds each at 60% power. Four sample tubes may be processed at once in a multi-tube rack which is positioned 1 to 3 mm above the opening in the probe. The DNA is returned to ice and a 1 µl sample is analyzed by electrophoresis on a 0.8% agarose gel in 0.5X TBE gel, run at 60 volts for 30 minutes. Sonication may be repeated if necessary.

A 0.26 µl aliquot of mung bean nuclease (150,000 u/ml) is added to sheared DNA and the sample is incubated at 30° C for 10 minutes. To stop the digestion, 20 µl of 1 M NaCl, 140 ul dd H20, and 200 ml of phenol:chloroform are added to the sample which is then, vortexed and centrifuged for 20 minutes at 13,000 rpm. The resulting aqueous phase is transferred into a new 1.5 ml microcentrifuge tube, 500 µl of 95% ethanol is added, and the DNA is precipitated overnight at -80° C. The sample is centrifuged for 30 minutes at 13,000 rpm, washed with 500 µl of 95% ethanol and centrifuged again for 30 minutes at 13,000rpm. The sample is then dried under vacuum, and resuspended in 10 µl TE.

The sheared DNA fragments are sized and purified by preparative agarose gel electrophoresis. Five microliters of 6x BP-XC-glycerol dye (0.25% BP, 0.25% XC, 30% glycerol) is added to the sample. The sample is split into two samples and loaded (12.5 µl per lane) on a 0.8% (1x TAE) low-melting agarose gel (SeaPlaque GTG) and electrophoresed at 60 V, 46 mA for 3.5 hours.

The gel is photographed under long wave UV and slices containing DNA fragments of 1.3 - 1.7 kb and 2 - 4 kb are excised and excess agarose cut away. The gel

slices are placed in 1.5 ml microcentrifuge tubes. One gel slice is stored at -20° C. 15 µl of 1 M NaCl is added to the other gel slice, followed by melting of the agarose by incubation at 65° C for 8 minutes. The resulting approximately 250 µl samples are placed into microcentrifuge tubes. An equal volume of water is added, following which the sample is vortexed and placed at room temperature for 2 minutes to bring the temperature up to 30 -35° C. 0.5 ml of water-saturated phenol that has been cooled on ice is added and the sample vortexed vigorously. The sample is placed on ice for 5 minutes, and the vortexing step repeated.

The sample is centrifuged at 4°C in a microcentrifuge for 20 minutes. The upper phase is transferred to a clean tube, and the bottom phenol layer is reextracted by addition of 200 µl of dd H2O. The sample is vortexed and placed on ice for 5 minutes, followed by centrifugation for 15 minutes. The aqueous layer is extracted and added to the aqueous layer from the previous step. Phenol extraction is repeated with 0.5 ml phenol, followed by vortexing and centrifugation for 20 minutes at 4°C. The aqueous layer is removed and repeated sec-butanol extractions are performed until the final volume is reduced to approximately 0.165 ml.

Two volumes of 95% ethanol (400 µl) are added and the sample is stored at -80° C overnight. The sample is centrifuged for 30 minutes at room temperature to pellet the DNA, washed once with 95% ethanol and dried briefly under vacuum. The sample is resuspended in 7 µl of TE. A 1 µl sample is run on a 0.8% agarose gel with markers to estimate concentration of recovered fraction.

<u>M13 Library</u>

20 ng of M13 DNA digested with *Sma*I is mixed with 1 µl of 10x ligation buffer (10X ligation buffer = 0.5M tris pH 7.4, 0.1M $MgCl_2$, 0.1M DDT), 1µl of 1mM ATP and 100 - 200 ng of sheared genomic DNA fragments (1 - 3 µl volume), and 0.3 µl of high concentration NEB ligase (5 unit/µl) is added. Water is added to a final volume of 10µl and the sample is incubated overnight at 14° C.

Plasmid Library

200 ng (4 µl) of pSTBlue vector (Novegene) is mixed with approximately 600 ng (12 µl) of sheared genomic DNA fragments from the 2-4kb size range gel slices and 1.2 µl of Gibco T4 ligase (5 units per µl) is added. Water is added to a final volume of 30µl and the sample is incubated overnight at 14° C.

Transformation

The ligation reaction is titered and diluted for optimal transformation efficiency. When the ligation contains approximately 20 ng of M13 vector, the dilution will typically be from 1:25 to 1:100. A 1:25 dilution is used for plasmid ligation containing approximately 200 ng of vector DNA. To increase transformation efficiency, the ligase is denatured by heating at 65°C for 7 minutes, and placed at room temperature for 5 minutes following the heating step.

A sterile electroporation cuvette is chilled for each transformation. Electro-competent cells are removed from the -80° C freezer and thawed on ice. For each M13 transformation, a sterile tube containing 25 ul of IPTG (25 mg/ml in water), 25 µl of X-Gal (25 mg/ml in dimethylformamide) and 3 ml of YT top agar is prepared, capped and placed in a 45° C water bath. YT plates are pre-warmed at 37° C for several hours to avoid cross-contamination problems that may result if water remains on plates. For plasmid transformations, a sterile tube containing 0.5 ml of SOC medium is prepared for each transformation, and L + amp plates are pre-spread with 25 µl of IPTG and 25 µl of X-Gal.

25 µl of electro-competent cells are mixed with DNA in diluted ligation mix in the cuvette, and the sample pulsed in an *E. coli* pulser (BioRad) set to the appropriate voltage (1.80kV for 0.1 cm cuvettes; 2.50kV for 0.2 cm cuvettes). The cuvette is removed from the pulser, and the sample immediately transferred to the tube containing SOC or YT top agar. For M13 transfections, the sample is plated immediately on YT plates. For plasmid transformations, the tube is placed in a 37° C shaker for 15-30

minutes and 30 ul aliquots are plated on L + Amp plates. Plates are incubated at 37° C overnight.

Example 2

Two basic methods can be used for DNA sequencing, the chain termination method of Sanger *et al., Proc. Natl. Acad. Sci. (U.S.A.) 74:*5463-5467 (1977), the entirety of which is herein incorporated by reference and the chemical degradation method of Maxam and Gilbert, *Proc. Natl. Acad. Sci. (U.S.A.) 74:*560-564 (1977), the entirety of which is herein incorporated by reference. Automation and advances in technology such as the replacement of radioisotopes with fluorescence-based sequencing have reduced the effort required to sequence DNA (Craxton, *Methods 2:*20-26 (1991), the entirety of which is herein incorporated by reference; Ju *et al., Proc. Natl. Acad. Sci. (U.S.A.) 92:*4347-4351 (1995), the entirety of which is herein incorporated by reference; Tabor and Richardson, *Proc. Natl. Acad. Sci. (U.S.A.) 92:*6339-6343 (1995), the entirety of which is herein incorporated by reference). Automated sequencers are available from, for example, Pharmacia Biotech, Inc., Piscataway, New Jersey (Pharmacia ALF), LI-COR, Inc., Lincoln, Nebraska (LI-COR 4,000) and Millipore, Bedford, Massachusetts (Millipore BaseStation).

In addition, advances in capillary gel electrophoresis have also reduced the effort required to sequence DNA and such advances provide a rapid high resolution approach for sequencing DNA samples (Swerdlow and Gesteland, *Nucleic Acids Res.. 18*:1415-1419 (1990); Smith, *Nature 349*:812-813 (1991); Luckey *et al., Methods Enzymol. 218*:154-172 (1993); Lu *et al., J. Chromatog. A. 680*:497-501 (1994); Carson *et al., Anal. Chem. 65*:3219-3226 (1993); Huang *et al., Anal. Chem. 64*:2149-2154 (1992); Kheterpal *et al., Electrophoresis 17*:1852-1859 (1996); Quesada and Zhang, *Electrophoresis 17*:1841-1851 (1996); Baba, *Yakugaku Zasshi 117*:265-281 (1997), all of which are herein incorporated by reference in their entirety).

A number of sequencing techniques are known in the art, including fluorescence-based sequencing methodologies. These methods have the detection, automation and instrumentation capability necessary for the analysis of large volumes of sequence data. Currently, the 377 DNA Sequencer (Perkin-Elmer Corp., Applied Biosystems Div., Foster City, CA) allows the most rapid electrophoresis and data collection. With these types of automated systems, fluorescent dye-labeled sequence reaction products are detected and data entered directly into the computer, producing a chromatogram that is subsequently viewed, stored, and analyzed using the corresponding software programs. These methods are known to those of skill in the art and have been described and reviewed (Birren *et al., Genome Analysis: Analyzing DNA*,1, Cold Spring Harbor, New York, the entirety of which is herein incorporated by reference).

PHRED is used to call the bases from the sequence trace files (http://www.mbt.washington.edu). Phred uses Fourier methods to examine the four base traces in the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations. Next, PHRED examines each trace to find the centers of the actual, or observed peaks and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations found in the first step.

After the base calling is completed, two sequence quality steps occur 1) poor quality end sequences are cut and if the resulting sequence is 50 bp or less it is deleted 2) overall sequence quality is examined and poor sequences are deleted from the data set if they have an average quality cutoff below 12.5. Contaminating sequences (*E. coli*, yeast, vector, linker) are removed after sequence quality assessment.

Contigs are assembled using PHRAP (http://www.mbt.washington.edu). Contigs and singletons are interrogated using AAT-NAP and BLASTP. AAT_NAP is a program used for constructing a global alignment of a DNA sequence and a protein sequence (Huang, X. *et al.*, *Genomics* 46:37-45 (1997), the entirety of which is herein incorporated by reference). The alignment model of NAP accommodates introns and frameshifts within codons. The scheme for scoring an alignment has several features that allow NAP to identify the exact locations of introns. A nucleotide insertion gap of length < k is given a linear penalty, and a nucleotide insertion gap of length > k is penalized as a gap of length k, where the value for k is the default value. The NAP program reports the starting and ending coordinates of predicted genes. The input to the NAP program includes the query sequence, the protein database and a coordinate file produced by AAT_EXT (an adapter between a database search program and a sequence alignment program) from the output of AAT_DPS (a program computing high-scoring chains of segment pairs between a query DNA sequence and the public non-redundant protein database from NCBI) The NAP program scans the protein database and finds the protein sequence for each coordinate record. Then for each coordinate record, NAP locates the query region, extends the region in both directions by a certain number of bases, and computes an alignment of the extended region and the protein sequence.

NAP corrects frameshifts in the query sequence.

BLASTP is used to validate the amino acid sequences and hits reported by the AAT_NAP program and to assign BLAST scores and p values to each sequence/hit pair. The AAT_NAP generated amino acid sequences are compared with the public non-redundant protein database (nr.aa from NCBI) using the default BLASTP parameters except that the V parameter is set to 1000000 (to report up to 1000000 hits that exceed the BLASTP default report cutoff). If the hit reported by AAT_NAP for a particular amino acid sequence is not reported by BLASTP, that particular amino acid sequence is removed.

Table 1

Seq Num	Contig Id	Gene Id	Position	Hit Id	AAT nap Score	Blast Score	Blast pvalue	%Ident	%Cvrg	Hit Description
1	ATC2C28183_1	ATC3On1	1093-1	g4309759	337	425	7.4e-40	84	20	(AC006217) unknown protein with Src homology 3 (SH3) domain profile (PDOC50002) [Arabidopsis thaliana]
2	ATC2C2_2237	ATC3On2	662-420	g4115361	122	161	4.9e-11	54	18	(AC005957) hypothetical protein [Arabidopsis thaliana]
3	ATC2C2_2234	ATC3On3	1-368	g4544372	469	495	9.4e-46	77	7	(AC006920) putative reverse transcriptase [Arabidopsis thaliana]
4	ATC2C2_2235	ATC3On4	446-854	g5541705	367	345	2.2e-31	51	66	(AL096860) putative protein [Arabidopsis thaliana]
5	ATC2C52466_1	ATC3On5	1-465	g4006824	391	500	8.3e-48	81	24	(AC005970) hypothetical protein [Arabidopsis thaliana]
6	ATC2C2_2236	ATC3On6	626-1	g4263644	443	547	8.7e-53	73	42	(AC006136) putative Athila-like protein [Arabidopsis thaliana]
7	ATC2C52466_2	ATC3On7	1-600	g4006824	409	489	1.2e-46	73	32	(AC005970) hypothetical protein [Arabidopsis thaliana]
8	ATC2C9421_1	ATC3On8	259-739	g1490554	773	720	4.0e-71	98	44	(U63633) S-adenosylmethionine decarboxylase [Arabidopsis thaliana]
9	ATC2C50782_1	ATC3On9	170-1	g4539387	291	300	1.3e-26	98	24	(AL035526) putative protein [Arabidopsis thaliana]
10	ATC2C2_2238	ATC3On10	1-510	g4773920	672	759	3.0e-75	83	23	(AF147259) F10A2.17 gene product [Arabidopsis thaliana]
10	ATC2C2_2238	ATC3On11	1-871	g4432802	324	416	8.0e-38	54	14	(AC006437) putative pol polyprotein, 5' partial [Arabidopsis thaliana]
11	ATC2C52466_3	ATC3On12	1-805	g4406797	837	940	2.0e-94	90	42	(AC006304) hypothetical protein [Arabidopsis thaliana]
12	ATC2C64710_1	ATC3On13	738-269	g4539440	210	263	8.1e-22	36	26	(AL049523) putative protein [Arabidopsis thaliana]
13	ATC2C23048_1	ATC3On14	1-933	g4006877	678	421	5.9e-55	89	51	(Z99707) RNA-binding like protein [Arabidopsis thaliana]
14	ATC2C21364_1	ATC3On15	668-1	g4454050	904	810	1.2e-80	94	59	(AL035394) putative protein [Arabidopsis thaliana]
15	ATC2C30011_1	ATC3On16	707-270	g3152576	180	225	7.7e-18	33	27	(AC002986) Similar to liver-specific transport protein gb|L27651 from Rattus norviegicus. [Arabidopsis thaliana]

Seq Num	Contig Id	Gene Id	Position	Hit Id	AAT nap Score	Blast Score	Blast pvalue	%Ident	%Cvrg	Hit Description
16	ATC2C2_2245	ATC3On17	1-339	g4309703	367	352	4.0e-32	68	43	(AC006266) putative transposon protein [Arabidopsis thaliana]
17	ATC2C8793_2	ATC3On18	859-1	g1694711	1039	1107	4.0e-112	98	31	(Y09581) FRO1 [Arabidopsis thaliana]
18	ATC2C64964_1	ATC3On19	1-458	g5541683	308	359	7.3e-33	43	45	(AL096859) receptor like protein kinase [Arabidopsis thaliana]
19	ATC2C2_2246	ATC3On20	61-613	g1723310	958	873	2.5e-87	98	68	hypothetical 30.2 KD protein in MODC-BIOA intergenic region [Escherichia coli]
20	ATC2C8793_1	ATC3On21	1-925	g2462833	1212	1180	7.3e-120	98	36	(AF000657) highly similar to froha and frohb, potential frohc [Arabidopsis thaliana]
21	ATC2C100_1	ATC3On22	1699-311	g5123928	2384	1742	2.0e-179	100	71	(AL079350) receptor kinase-like protein [Arabidopsis thaliana]
22	ATC2C51584_1	ATC3On23	654-1	g3935184	430	535	2.0e-50	90	10	(AC004557) F17L21.27 [Arabidopsis thaliana]
23	ATC2C19508_1	ATC3On24	1-1252	g4580469	273	352	4.0e-32	38	56	(AC006081) putative zinc finger protein [Arabidopsis thaliana]
24	ATC2C2602_1	ATC3On25	246-1	g3927836	117	171	7.7e-12	42	13	(AC005727) unknown protein [Arabidopsis thaliana]
25	ATC2C2_2252	ATC3On26	432-132	g2495537	495	499	1.1e-47	92	43	hypothetical 25.5 KD protein in HUPB-COF intergenic region [Escherichia coli]
26	ATC2C2_2248	ATC3On27	587-82	g4335720	678	693	2.5e-67	78	14	(AC006248) putative reverse transcriptase Tal-1 [Arabidopsis thaliana]
27	ATC2S60010_1	ATC3On28	520-1	g99719	255	205	6.1e-18	39	16	hypothetical protein 2 - Arabidopsis thaliana retrotransposon Ta1-2 (strain Landsberg) (fragment) [Arabidopsis thaliana]
28	ATC2C2_2251	ATC3On29	1-687	g4263831	855	941	8.8e-94	77	16	(AC006067) putative reverse transcriptase [Arabidopsis thaliana]
29	ATC2C2_2250	ATC3On30	759-1	g2129618	728	604	2.8e-58	61	27	hypothetical protein 1 - Arabidopsis thaliana retrotransposon Athila [Arabidopsis thaliana]
30	ATC2C12463_1	ATC3On31	1191-1	g2244904	203	278	2.8e-23	27	37	(Z97339) hypothetical protein [Arabidopsis thaliana]
31	ATC2C2_2258	ATC3On32	1-728	g4455297	458	479	1.4e-45	60	55	(AL035528) hypothetical protein [Arabidopsis thaliana]
32	ATC2C2_2257	ATC3On33	629-1	g3319757	83	124	1.2e-06	31	17	(AL031035) putative ATP /GTP-binding protein [Streptomyces coelicolor]
33	ATC2C43311_1	ATC3On34	1-857	g3461848	1177	1156	2.5e-117	94	31	(AC005315) putative ATPase [Arabidopsis thaliana]

Seq Num

A sequential SEQ ID NO: is assigned to each contig or singleton and the SEQ ID NO: corresponds to that set forth in the sequence listing.

Contig Id

Contigs or singletons are assigned an arbitrary contig ID. Contigs are assembled according to the procedure set forth in Example 2.

Gene Id

Refers to an arbitrarily assigned Gene ID number.

Position

If the first numeral under the position heading is lower than the second numeral, it designates the nucleotide position which forms part of the codon that encodes the N-most terminal amino acid of the coding sequence of the *Arabidopsis* protein or fragment thereof. If the first numeral under the position heading is higher than that found in the corresponding second position it designates the nucleotide position which forms part of the codon that encodes the C-most terminal amino acid of the coding sequence of the *Arabidopsis* protein or fragment thereof. In cases where the first numeral is higher than its corresponding second numeral, the *Arabidopsis* protein or fragment thereof is encoded by the complement of the sequence set forth in the sequence listing.

Hit Id

Each sequence in the GenBank public database is arbitrarily assigned a unique Hit Id or (National Center for Biotechnology Information GenBank Identifier) NCBI gi number. In this table, the Hit Id or NCBI gi number which is associated (in the same row) with a given contig or singleton refers to the particular GenBank sequence which is the best match for that sequence.

AAT nap score

The AAT nap score is reported by the nap program in the aat package. It is an alignment score in which each match and mismatch is scored based on the BLOSUM62 scoring matrix.

Blast Score

Each entry in the "Blast Score" column of the table refers to the BLASTP score that is generated by sequence comparison of the designated clone with the designated GenBank sequence.

Blast pvalue

The entries in the "Blast pvalue" column refer to the probability that such matches occur by chance.

%Ident

The entries in the "%Ident" column of the table refer to the percentage of identically matched nucleotides (or residues) that exist along the length of that portion of the sequences which is aligned by the BLAST comparison to generate the statistical scores presented.

%Cvrg

The % coverage is the percent of hit sequence length that matches to the query sequence (% coverage = (match length / hit total length) x 100).

Hit Description

The "Hit Description" column provides a description of the NCBI gi referenced in the "Hit Id" column.